British Journal of Educational Technology *Vol 33 No 5 2002* 593–602

Paper-based versus computer-based assessment: key factors associated with the test mode effect

Roy Clariana and Patricia Wallace

Roy B Clariana is at the School of Graduate Professional Studies, The Pennsylvania State University and Patricia E Wallace is at the School of Business, The College of New Jersey. Address for correspondence: Roy B Clariana, Great Valley School of Graduate Professional Studies, The Pennsylvania State University, Malvern, PA 19355, USA. Tel: +1 610 648 3253; email: RClariana@psu.edu

Abstract

This investigation seeks to confirm several key factors in computer-based versus paper-based assessment. Based on earlier research, the factors considered here include content familiarity, computer familiarity, competitiveness, and gender. Following classroom instruction, freshman business undergraduates (N = 105) were randomly assigned to either a computer-based or identical paper-based test. ANOVA of test data showed that the computer-based test group outperformed the paper-based test group. Gender, competitiveness, and computer familiarity were NOT related to this performance difference, though content familiarity was. Higher-attaining students benefited most from computer-based assessment relative to higher-attaining students under paper-based testing. With the current increase in computer-based assessment, instructors and institutions must be aware of and plan for possible test mode effects.

As pointed out by other authors in this special issue, for many reasons, the use of computer-based assessment is increasing. Examples include state drivers' license exams, military training exams, job application exams in the private sector, entrance exams in postsecondary education, and certification exams by professional groups (Russo, 2002; Trotter, 2001). However, in the literature, there is mounting empirical evidence that identical paper-based and computer-based tests will not obtain the same results. Such findings are referred to as the "test mode effect."

For example, paper-based test scores were greater than computer-based test scores for both mathematics and English CLEP tests (Mazzeo, Druesne, Raffeld, Checketts and Muhlstein, 1991) and for recognizing fighter plane silhouettes (Federico, 1989); while computer-based test scores were greater than paper-based test scores for a dental hygiene course unit midterm examination (DeAngelis, 2000); though other studies

Published by Blackwell Publishers Ltd, 108 Cowley Road, Oxford, OX4 1JF, UK and 350 Main Street, Malden, MA 02148, USA.

have reported no difference between computer and paper-based tests (Mason, Patry and Berstein, 2001; Schaeffer, Reese, Steffen, McKinley and Mills, 1993).

How common is the test mode effect? In a review of educational measurement approaches, Bunderson *et al* (1989) reported three studies that showed a superiority for computer-based tests, eleven studies that showed no difference, and nine studies that showed a superiority for paper-based tests. Based on their findings, the chances of any particular test obtaining equivalent results on paper and computer are only about 50%.

How much different are computer-based versus paper-based test scores? Bunderson *et al* (1989) state, "... the scores on tests administered on paper were more often higher than on computer-administered tests... the score differences were generally quite small..." (p. 378). Mead and Drasgow (1993) in a meta-analysis of computer versus paper-based cognitive ability tests also found that on average, paper-based test scores were very slightly greater than computer-based test scores. Note that, though the difference may be small, the consequences for an individual student may be substantial (ie, pass versus fail).

Instructional design dogma insists that paper-based versus computer-mediated instructional components should produce exactly equivalent results if the content and cognitive activities of the two are identical (Clark, 1994). In most test mode effect studies, the computer-based and paper-based versions are nearly identical and the cognitive activity required to answer a test item on paper or computer should be the same, yet significant differences are regularly observed (Bunderson, Inouye and Olsen, 1989). Paraphrasing Clark, though the kind of truck used to deliver groceries cannot impact the nutrition of the groceries, if the learner has never driven a truck, the groceries may not be delivered at all. Examining individual characteristics of learners provides one promising avenue for determining the key elements involved in the test mode effect.

First, individual learner characteristics have been used to account for differences between computer-based and traditional instruction (Wallace and Clariana, 2000; Watson, 2001). For example, Wallace and Clariana (2000) completed a program evaluation of a four-week long spreadsheet module converted to online delivery. Learner characteristics associated with higher posttest performance for the web-based group (relative to the traditional face-to-face group) included content familiarity, computer familiarity, and non-competitiveness. The authors concluded that learners who were less familiar with the content, who were less familiar with computers, and who are competitive would not do as well online, and so these learners should be allowed or even required to take that module in the traditional classroom setting. Similarly, Watson (2001) reported that students with higher academic attainment and also those with greater frequency of computer use benefited most from computer-aided learning.

More on point, some learner characteristics have been directly associated with the test mode effect. In a study of the Graduate Record Examination (GRE) delivered by computer and paper, Parshall and Kromrey (1993) reported that computer-based test scores on the verbal, quantitative, and analytic sections of that test were all greater than complementary paper-based test scores. Here, gender, race, and age were associated with test mode. In general, white males did better with computer-based delivery while males in other racial groups did best with the paper-based tests, though there was no difference for females for paper or computer-based tests. They also reported that neither preference for computer or paper-based tests, nor prior experience with computers were associated with test mode differences. However, note that this population of examinees was highly computer literate and self-selected to take the computer-based version.

Based on these previous findings, this investigation compares computer-based and paper-based test scores of a 100-item teacher made multiple-choice test of facts and concepts related to general knowledge of computers that was covered in class lectures and course readings. Post-test data and learner self-report information were analyzed in order to confirm some key factors that relate to the test mode effect. Based on earlier research, the key learner characteristics considered here include prior content familiarity, computer familiarity, competitiveness, and gender (Parshall and Kromrey, 1993; Wallace and Clariana, 2000; Weaver and Raptis, 2001).

Methodology

Design and sample

This study used a post-test only design with one factor, test mode (computer-based and paper-based). Dependent variables included students' scores on a 100-item multiple choice test and also students' self-report on a distance learning survey.

Four sections of the course, Computer Fundamentals, consisting of 105 students were selected as the sample for this investigation. Two sections consisting of 51 students were randomly selected as the paper-based test group. Two other sections consisting of 54 students were identified as the computer-based test group.

The course and procedure

Business 100, Computer Fundamentals, is the first computer course in the Business School that students are advised to complete in their freshman year. This course covers the fundamental concepts and use of a computer system. The course is divided into three units of instruction: an introduction to the Windows operating environment, the use of electronic communications, and financial applications using a spreadsheet software package. Course assignments involve Internet searching, word processing, designing a web page, and using a spreadsheet.

Early in the course, to establish an overview, course vocabulary, and requisite fundamental knowledge, general computer-related facts and concepts are covered in textbook readings and lectures and then student fundamental knowledge is tested. These

facts and concepts cover such topics as computer storage devices, types of printers, resolution of monitors, computer peripheral devices, types of modems, and so on. The test of this unit of the course is the dependent variable in this investigation.

This fundamental knowledge was covered early in the course and students were tested as soon as the material was covered. The course syllabus handed out to the students on the first day of class listed this up-coming test and that it was worth 15% of their final course grade. On test day, students first completed a self-report distance learning profile and then completed the test either on computer or on paper.

Post-test and self-report instruments

The test consisted of 100 multiple-choice questions each with four alternatives. On the paper-based version, six or seven questions were written on each page. Students read each question and then wrote the letter (A, B, C, or D) of the answer choice on a separate answer sheet. With the computer version, students received one question per screen. Students clicked on the letter of the correct answer choice and then proceeded to the next question. Students could review and change previously answered questions. The item order was not the same on the paper and computer versions; on the computer version, the test questions were randomized.

To measure examinee characteristics, the Distance Learning Profile (DLP) was used (Clariana and Moller, 2000). The DLP consists of 24 statements that have been shown to strongly relate to distance learning course performance. These statements fall into five categories (factors) including: Active Engagement, Independence, Competitiveness, Perceived Course Quality, and Frequency of Computer use.

Results

Student achievement comparisons

A one-factor between-subjects ANOVA was computed to compare the computer and paper-based groups' test means. The ANOVA obtained an $F(1,103) = 15.324$, $MSE = 80.363$, $p < 0.001$, which indicates a large statistically significant difference between the scores of the computer-based testing group (83.0, sd = 8.7) compared to the scores of the paper-based testing group (76.2, sd = 9.3).

Test mode and examinee characteristics

Several separate planned 2-factor between-subjects ANOVAs related to examinee characteristics were conducted. The characteristics considered in these separate analyses as previously described included: gender (coded male and female), content familiarity, computer familiarity (coded high and low based on DLP items 21, 22, 23, and 24), and competitiveness (coded low and high based on combined DLP items 6, 13, and 18; from Clariana and Moller, 2000). In this case, gender, computer familiarity, and competitiveness were not associated with computer versus paper test mode effects, however content familiarity was. A 2x2 ANOVA with the factors test mode (paper and computer) and content familiarity (coded low and high based on median split of final course grade) was conducted. As above, a significant difference for test mode was observed, with

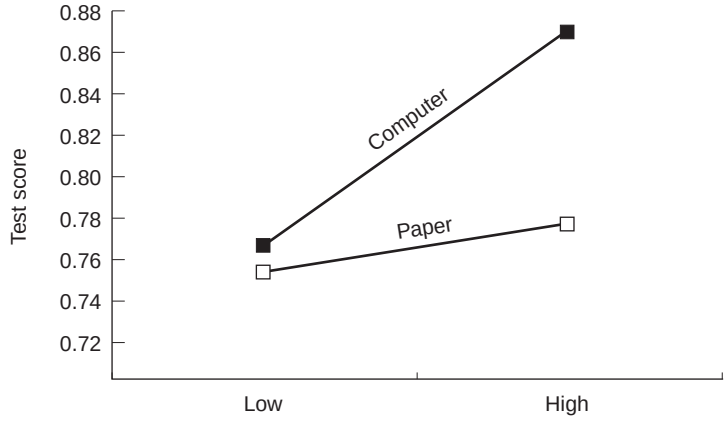

Figure 1: Interaction of test mode (computer vs. paper) and content attainment (low vs. high)

F(1,101) = 9.641, MSE = 69.973, p = 0.002. The content familiarity factor was also significant, F(1,101) = 12.483, MSE = 69.973, p = 0.001, with high-attaining students obviously outscoring low-attaining students. The interaction of test mode and content familiarity was also significant F(1,101) = 5.066, MSE = 69.973, p = 0.027 (see Figure 1). Computer-based testing especially helped the high-attaining students (relative to paper-based testing), or alternately, paper-based test administration hindered the performance of high-attaining students.

Distance learning profile predictors

To determine whether any of the DLP items directly predict course achievement, simple correlations of each DLP item with post-test scores were conducted. On the paper-based test, six DLP items were significantly related to post-test achievement (see Table 1). Specifically, students who reported that the course and activities were boring (items 4 and 10), who reported that they are competitive (items 6 and 9), and those who really don't care how others in class are doing (item 14) performed best on the paper-based test. We will use the term "egocentric" to describe this non-collaborative, adverse set of qualities. Thus, the egocentric students scored highest on the paper-based posttest. On the computer-based test, only item 13 was related to achievement. Specifically, students who reported that they work harder than others performed best on the computer-based test.

Principle components factor analysis was conducted with the Distance Learning Profile (DLP) data to establish factor item groups for the current investigation. Four factors were identified that were similar to the five factors described in the previous study (Clariana and Moller, 2000). These include: *Active Engagement* (DLP# 4, 10, 16, 17), *Course Engagement* (1, 3, 20, 21, 22), *Independence* (2, 5, 11, 14), and *Competitiveness* (6, 13, 18, 23, 24). The DLP items designed to measure computer familiarity (items 21,

Table 1: Correlation of DLP and test performance

P (r)	*C (r)*	*Factor*	*DLP Item*
–0.18	–0.02	CE	1. Course assignments are interesting.
0.21	0.17	IN	2. I learn best without supervision.
–0.10	0.17	CE	3. I prefer tough courses that really challenge me.
0.40	–0.03	AE	4. Many of the course activities seem useless.
0.21	0.22	IN	5. I am a self-starter.
0.33	0.18	CP	6. I always try to out perform other students.
0.04	0.07	–	7. The course assignments are appropriate.
–0.12	0.02	–	8. I usually prepare for exams well in advance.
–0.24	0.02	CP	9. I make sure that other students get my viewpoint.
0.29	–0.06	AE	10. The course is boring.
–0.18	–0.14	IN	11. I prefer constant feedback from the teacher.
0.24	–0.13	–	12. My views contribute little to the quality of a course.
0.01	**0.29**	CP	13. I work harder than others to stand out from the crowd.
0.50	0.02	IN	14. I don't care how others are doing on assignments.
0.22	0.14	–	15. I work best under a deadline.
–0.03	–0.07	AE	16. This course actively engages me.
–0.06	–0.01	AE	17. Overall, I consider this to be a high quality course.
0.10	0.00	CP	18. I am usually competitive.
–0.05	–0.18	–	19. I prefer to do assignments my way.
–0.05	0.03	CE	20. This course "turns me on".
0.14	0.21	CE	21. I use computers everyday.
0.02	0.19	CE	22. I often use the Internet.
0.03	–0.16	CP	23. I don't like computers.
0.11	0.03	CP	24. I often access my e-mail.

P—Paper, C—Computer; Factors—Active Engagement (AE), Course Engagement (CE), Independence (IN), and Competitiveness (CP). (Note: r values significant at $p < 0.05$ are shown in bold)

22, 23, and 24) grouped here with Course Engagement and Competitiveness items. This new grouping of computer use items likely occurred because of the heavy emphasis on computers in this course.

Discussion

In this investigation, computer-based test delivery positively impacted scores on this instructor-made test of course content relative to paper-based testing. This was especially apparent with higher-attaining students. Though computer familiarity, gender, and competitiveness are learner characteristics that have previously been related to test mode effects, in this present investigation, none were related to the test mode effect, however, content familiarity was related to the test mode effect. Specifically, computer-based tests especially helped the high-attaining students (relative to paper-based testing).

Further, several interesting correlations with achievement were observed. The DLP items here obtained strikingly different correlations with paper-based versus computer-based test performance. The correlation results (see Table 1 again) suggest that those

students who approached the paper-based tests with an egocentric stance did best on the post-test (relative to the less egocentric students). They seem to be directly competing with each other for grades, and this stance in some way relatively improved their test performance. This relationship between competitiveness and post-test achievement was also observed with the computer-based test group, but was far less of a factor. Wallace and Clariana (2000) also reported that competitiveness was significantly related to achievement for a traditional face-to-face instruction group, but competitiveness was not related to achievement for the parallel online group. In both cases, perhaps the novelty of computer delivery mitigated taking an egocentric stance. If so, as students become as familiar with computer-based testing as they are with paper-based testing, the test mode effect should decrease or disappear. In any case, we recommend that future investigations of the test mode effect include some measure of computer familiarity, content familiarity (or general content attainment), engagement with the course, competitiveness, and also a measure of previous computer-based testing should be collected.

In any investigation of the test mode effect, the exact equivalency of the paper and computer forms must be closely examined. Though the computer-based and paper-based test versions here were identically worded and both allowed the students to review and change their responses, there are possible differences between the two. For example, Mourant, Lakshmanan and Chantadisai (1981) have shown that students become more fatigued when reading text on a computer screen than when they read the same text on paper. Also, different fonts (typefaces appear differently on computer and paper) have been related to computer versus paper differences (Wilson, 2001). Thus, previous findings for paper-based over computer-based assessment performance may relate in some way to the poor resolution of computer monitors (which also impacts the readability of the fonts used). However, screen resolution now is much better than previously. Recall that here, the computer test group outperformed the paper test group, thus any negative effects of poor computer screen resolution were not evident in this investigation. If screen resolution is a factor in the test mode effect, its affects should continually decrease as screen resolution improves.

Next, item order (computer administered test items are often presented in a randomized order) and the order of multiple-choice response options (sometimes randomized in computer administered tests) can affect performance on an item (Beaton and Zwick, 1990; Cizek, 1991). This likely relates to "ordered" versus randomized test item sequencing. Specifically, when the instructional lesson content and the test items are in the same order, the "ordered" test will likely obtain greater scores than a randomized version of the test. In the present investigation, both the paper version and the computer versions of the test were randomly generated, thus mitigating an order effect.

Probably the two greatest physical differences between computer and paper test administration involve perceived interactivity and physical size of the display area. The amount of information comfortably presented in a computer display is only about one-third of that presented by a standard piece of paper. For example, Haas and Hayes

(1986) reported that when a text passage associated with a test item requires more than one page, computer administration yielded lower scores than paper-and-pencil administration, apparently due to the difficulty of reading the extended text on screen. In this case text passages were not used, however, on the paper-based test, several test items were presented on each piece of paper. The student can rapidly scan all of the questions on a page and can easily flip backward or forward to other pages (a form of interactivity). On the computer-based assessment, one test item was presented on each computer screen display and the student must physically act to move from screen (item) to screen (another form of interactivity). This difference likely leads to greater "focus" and closure with each computer-based item. Thus computer-based items (relative to paper) may increase transition time and memory load, with a tighter focus on and closure of each individual item (Clariana, 1997; Clariana and Smith, 1988).

Both high- and low-able students should benefit from greater focus on an item; though due to the greater cognitive load required, only high-able students would be able to tie ongoing items together to "learn" from the test in order to answer other test items. To examine this hypothesis, a test could be designed that intentionally provides items that, if remembered, will allow the student to answer other items correctly. If high-able learners do learn during the test (relatively), a pattern of means similar to that observed in this present investigation should occur. Also, in order to parallel computer test displays, paper-based tests could be designed with only one question per page and this test format can be compared to a paper-based test with multiple-questions per page. If display size format is the primary factor, then the multiple-page group should out perform the one-item per page format.

Another possible explanation of the test mode effect is transfer appropriate processing (TAP; Bransford and Franks, 1976; Morris, Bransford and Franks, 1977). TAP proposes that lesson and assessment processes should correspond. Since the unit of instruction involved learning about and using computers including extensive hands-on computer activities every class period, then the computer test mode may better "match" the lesson approach than does the paper test mode. To determine the possible role of TAP in the test mode effect, a future study should completely cross an "identical" paper-based and computer-based lesson (problematic from an instructional design viewpoint) with identical paper-based and computer-based assessment. If the "paper" lesson is better with the "paper" test and the "computer" lesson with the "computer" test, then a TAP explanation would be supported.

In summary, establishing a model that fully accounts for test performance differences for computer versus traditional examination approaches may be some time away, however, because of the growth of computer-based testing, it seems critical at this time to further this line of research. Based on our review and these results, we anticipate that computer familiarity is the most fundamental key factor in the test mode effect, especially for unfamiliar content and/or for low attaining examinees (especially an issue for students with reduced computer access, such as women and minorities). In general, higher-attaining students will adapt most quickly to any new assessment

approach (Watson, 2001) and will quickly develop test-taking strategies that benefit from the new approach. Thus, in the current investigation, as students learned about computers, the higher-attaining students likely accommodated more quickly and so benefited more from computer-based assessment. Once all students are fully familiar with computers, then computer familiarity should become less important, though other factors associated with traditional testing are likely to emerge in computer-based testing, such as competitiveness, need for achievement, and independence, as well as new forms of "cheating".

Regardless of the key factors at work in the test mode effect, the findings of this study have rather practical implications for the students completing this course. In this case, the computer-based test group mean was 83%, a low B grade, and the paper-based test group mean was 76%, a C grade. At a minimum, the paper-based test scores should be adjusted upward in line with computer-based scores or vice versa, otherwise it would seem to be unfair. On the other hand, which of these assessment modes more accurately reveals the students' actual knowledge? Future research should include complementary assessment measures of the same content in order to establish the criterion-related evidence for computer-based and paper-based approaches.

Finally, Bugbee (1996) recommends that test developers must show that computer-based and paper-based test versions are equivalent, and/or must provide scaling information to allow the two to be equated. Most instructors, and in fact, even most instructional designers, do not have the skill nor the time to craft and extensively pilot their examinations. However, additional time and effort must be invested by instructors to design high-quality test items for use in online testing. With the likely proliferation of web-based courses and of inexpensive fingerprint identification computer devices and other automatic proctoring technologies, there will likely be a substantial increase in computer-based testing. The findings of this investigation indicate that it is critical to realize that computer-based and paper-based tests, even with identical items, will not necessarily produce equivalent measures of student learning. Instructors and institutions should spend the time, cost, and effort to mitigate test mode effects.

References

Beaton A E and Zwick R (1990) *The effect of changes in the National Assessment: disentangling the NAEP 1985–86 reading anomaly* Educational Testing Service, Princeton, NJ.

Bransford J D and Franks J J (1976) *The role of "effort after meaning" and "click of comprehension" in recall of sentences* Educational Resources Document Reproduction Service (ERIC) # ED188208.

Bugbee A C Jr (1996) The equivalence of paper-and-pencil and computer-based testing. *Journal of Research on Computing in Education* **28** (3) 282–299.

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.

Cizek G J (1991) *The effect of altering the position of options in a multiple-choice examination* A paper presented at the Annual Meeting of the National Council on Measurement in Education in Chicago, IL, April 4–6 (Educational Resources Document Reproduction Service (ERIC) # ED333024).

Clariana R B (1997) Considering learning style in computer-assisted learning *British Journal of Educational Technology* **28** (1) 66–68.
Clariana R B and Moller L (2000) *Distance learning profile instrument: predicting on-line course achievement* Presented at the Annual Convention of the Association for Educational Communications and Technology, Denver, CO. [http://www.personal.psu.edu/rbc4/dlp_aect.htm]
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 Computers in Education (IACE), New Orleans, LA, April, 1988 (ERIC Document Reproduction Service: ED 295 796).
Clark R E (1994) Media will never influence learning *Educational Technology Research and Development* **42** (2) 21–29.
DeAngelis S (2000) Equivalency of computer-based and paper-and-pencil testing *Journal of Allied Health* **29** (3) 161–164.
Federico P A (1989) Computer-based and paper-based measurement of recognition performance *Navy Personnel Research and Development Center Report NPRDC-TR-89-7* (Educational Resources Document Reproduction Service (ERIC) # ED306308).
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.
Mason B J, Patry M and Berstein D J (2001) An Examination of the equivalence between non-adaptive computer-based and traditional testing *Journal of Educational Computing Research* **24** (1) 29–39.
Mazzeo J, Druesne B, Raffeld P C, Checketts K T and Muhlstein A (1991) Comparability of computer and paper-and-pencil scores for two CLEP general examinations *College Board report No. 91-5* (Educational Resources Document Reproduction Service (ERIC) # ED344902).
Mead A D and Drasgow F (1993) Equivalence of computerized and paper-and-pencil cognitive ability tests: a meta-analysis *Psychological Bulletin* **114**, 449–458.
Morris C D, Bransford J D and Franks J J (1977) Levels of processing versus transfer appropriate processing *Journal of Verbal Learning and Verbal Behavior* **16**, 519–533.
Mourant R R, Lakshmanan R and Chantadisai R (1981) Visual fatigue and cathode ray tube display terminals *Human Factors* **23** (5) 529–540.
Parshall C G and Kromrey J D (1993) *Computer testing versus paper-and-pencil: an analysis of examinee characteristics associated with mode effect* A paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA, April (Educational Resources Document Reproduction Service (ERIC) # ED363272).
Russell M (1999) Testing on computers: a follow-up study comparing performance on computer and on paper *Education Policy Analysis Archives* **7**, 20. Available online: http://epaa.asu.edu/epaa and also at http://www.ed.gov/Technology/TechCerf/1999.
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
Schaeffer G A, Reese C M, Steffen M, McKinley R L and Mills C N (1993) Field test of a computer-based GRE general test *ETS Research Report #93-07* (Educational Resources Document Reproduction Service (ERIC) # ED385588).
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** (1) 13–18. [http://gise.org/JISE/Vol11/v11n1-2p13-18.pdf]
Watson B (2001) Key factors affecting conceptual gains from CAL *British Journal of Educational Technology* **32** (5) 587–593.
Weaver A J and Raptis H (2001) Gender differences in introductory atmospheric and oceanic science exams: multiple-choice versus constructed response questions *Journal of Science Education and Technology* **10** (2) 115–127.
Wilson R F (2001) *HTML E-mail: text font readability study*. Results of a survey conducted April, 2001. Available at: http://www.wilsonweb.com/wmt6/html-email-fonts.htm