# Reading and Searching Digital Documents: An Experimental Analysis of the Effects of Image Quality on User Performance and Perceived Effort

**Andrew Dillon, Lisa Kleinman, Randolph Bias, Gil Ok Choi & Don Turnbull**

*School of Information, University of Texas at Austin, 1 University Station, D7000, Austin, TX 78712.*
*Email: {adillon, kleinman, rbias, gilok, donturn}@ischool.utexas.edu*

**As increasing amounts of information are viewed, read and manipulated digitally, many users still report performance and satisfaction costs with digital documents compared to paper equivalents. While many factors impact this process, image quality has been assumed by many to have relatively little impact once current screen display standards are maintained. We test this hypothesis by comparing users on a variety of routine information tasks performed on standard and enhanced screen displays. Using Microsoft ClearType, a font technology which enhances text display resolution by accessing vertical color stripes at the pixel level, we examined user performance on reading, editing and searching tasks with routine office applications. Results suggest that for tasks involving lengthy eye-on-text interaction (e.g., reading for comprehension) advances in image quality can still yield significant improvements in performance for most users.**

## Introduction

Understanding and influencing the process of reading digital documents has led to a substantial empirical research literature on reading from screens versus reading from paper. This research indicates that sources of difference between these media can cause electronic text to be read up to 25% slower and, removing all effects associated with variables such as navigation, manipulation, portability and preference, the quality of image presented on screen can prove a significantly limiting factor.

Originally identified as such by Gould et al. (1987), subsequent work by Muter and Mauretto (1991) confirmed that the quality of the image presented on screen could be improved to a point where differences between paper and screen reading, at least in terms of reading speed, could be reduced to non-significant levels. Consequently, as screen technologies have improved, research into image quality has taken a back seat to studies of user navigation in large digital documents and willingness to read lengthy texts online (see e.g., O'Hara and Sellen, 1997).

Image quality may yet be worthy of further attention from researchers, not least because the trend towards increased reading of documents online, across a variety of screen sizes and styles (PDAs, laptops, desktops, mobile phones etc.) means that comparisons with paper are no longer the sole measure of success. With billions of dollars of work time spent each day across multiple organizations by people reading electronic texts, any improvements in speed that are extracted from screen display technologies could have very significant financial and performance effects.

Microsoft's ClearType technology promises significant improvement in the visual quality of text presentation on screen. ClearType is a setting that is manipulated through the Windows operating system of devices using Liquid Crystal Displays (LCDs), and works by altering the vertical color stripe within a pixel allowing for changes in how the text looks at fractional levels. These changes aim to enhance the resolution of the screen text and thus improve readability. While font design has continued to be studied as a factor in electronic document presentation (see e.g., Boyarski et al., 1998, Bernard et al., 2003), ClearType can work *across* multiple fonts and may have more applicability in a world where many users like to choose their own font displays.

The original image quality work of the late 1980s and 1990s concentrated on highly controlled tasks, such as proofreading, which was often chosen for its ease of control and assessment of performance. However, most human reading is far less structured, involving browsing, searching, reading for meaning or entertainment, or reading for gist. Dillon (2004) posits that reading is best understood in user experience terms as a combination of image processing, document manipulation, and modeling of structure in pursuit of a task goal in a given context. Accordingly, there is no single variable that will explain the performance or preference outcomes since the user's response is determined by multiple factors.

Any improvement in image quality will likely have maximum impact on the phases of the reading process where the eyes meet the text. Decomposing the reading task into these component phases can guide evaluation and research by suggesting what proportion of the task will benefit. For tasks consisting of large proportions of serial visual processing of text (e.g., serial reading of multiple paragraphs), then image quality is likely to be a major determinant of reading speed or efficiency. For tasks where the user jumps, manipulates, or otherwise removes the eyes from continuous physical engagement with the words (e.g., navigating a web site), then image quality will likely prove less important to performance and other factors such as information layout and organization will become more influential. Hence, our hypothesis that effects of ClearType will be highly task-dependent.

## Overview of Study

The present study is a first attempt to test the task-dependency qualification of the image quality hypothesis. Users are given three distinct tasks in a block (editing a marked up manuscript, scanning and extracting data from a spreadsheet display, and reading an article for comprehension) over two computer display conditions (ClearType-on and ClearType-off, hereafter referred to as ClearType and Regular text).

These tasks were designed to require a range of physical, perceptual and cognitive acts by users. For the editing tasks, users were given a marked up paper manuscript and its original MS Word version (for an example screen see Figure 1). From here participants had to make the required edits on the electronic version. This task required multiple shifts of gaze from paper to screen and the identification on screen of specific locations. Typical edits involved changing a number or a short sequence of letters. This task was considered a good index of the impact of ClearType on short duration scanning and visual location.

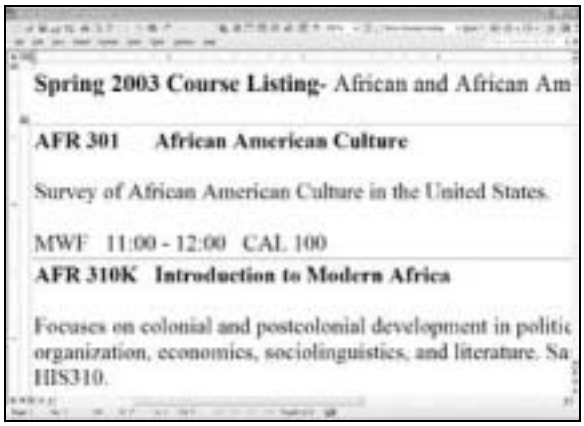



Figure 1: Example screen for the editing task

For the scanning tasks, users examined a staffing schedule presented in MS Excel where codes were used to name individual staff members. The task involved locating times and days when specific combinations of staff members were on duty. This was also a scanning task but one that involved greater eye-on-screen activity as the user searched the spreadsheet for combinations of staff and linked these to heading information on time and day. A partial example of the type of information displayed for this task is presented in Figure 2.

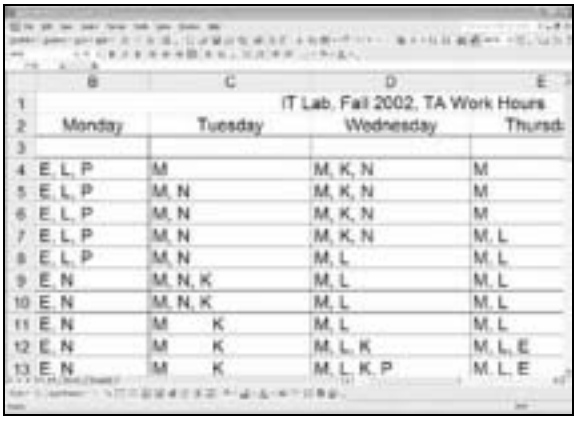



Figure 2: Example screen for the scanning task

Finally, the reading tasks involved users reading a 2000 word article on a topic of general interest presented as a single scrollable text in a Mozilla Firebird browser window. To ensure reading occurred, participants were told they would be asked several questions upon completion. This task involved consistent visual processing of the text with occasional scrolling to display the full document, a typical form of screen reading for most users. Participants did not take notes, nor were they able to refer back to the article while answering the questions.

In this way we obtained data on use of electronic documents across a range of likely scenarios involving repeated or consistent visual perception, occasional physical manipulations, visual search and cognitive processing activities. To supplement the performance measures, the NASA Task Load Index (NASA-TLX) was used between tasks to capture user estimates of cognitive effort and fatigue. The TLX presents users with six scales covering issues such as effort and demand perceived by the participant in the course of completing the task. Answers are expressed on a 20-point scale. TLX is considered the standard tool for measuring cognitive effort. It was employed as a means of capturing any perceived benefits for ClearType beyond behavioral measures.

## Method

### *Participants*

Participants were recruited through posted advertisements at the University of Texas at Austin and online job web sites. All participants met the following criteria:

- age 18 or older,
- English as a first language,
- self-reported familiarity using MS Word, MS Excel, and Internet Explorer,
- self-reported as having 20/20, standard, or corrected vision,
- self-reported as having no reading disabilities or color blindness.

38 participants were scheduled for 1-hour time slots and paid $10 for their involvement.

### *Experimental Design*

The study employed a within-subjects (repeated measures) design with each participant completing three task types in two blocks of trials, for a total of six tasks. The independent variable was the manipulation of text display across these tasks operationalized at the following levels:

- display setting (ClearType, Regular)
- task type (editing, scanning, reading)
- content type (content A, content B)

Performance was assessed through three dependent variables:

1. time to complete each task
2. accuracy of task answers
3. NASA-TLX responses for each task

Test software was developed to enable automatic capture of time data for each task. It started running at the moment the task was visible on the computer screen and ended when the participant clicked a “Done” button. Participants controlled when they began and finished each task.

Task accuracy data was calculated for each task. For the editing task, participants made 16 edits in a MS Word document based on a provided paper-version of the document with red-ink corrections. Accuracy scores reflect the number of correct edits made in the electronic version. In the scanning task, participants answered three questions about the spreadsheet information, and each answer was either correct or not. Two of these scanning questions were multi-part, making for a total possible 5 correct answers. For the reading task the participant was asked to respond to two multiple-choice questions, one which directly asked for factual information stated in the article, and one which required inference based on the material they had read. Additionally, the participant was asked to summarize the article in three to five sentences, and a total word count of these answers was calculated. The NASA-TLX responses were collected after each task and consisted of seven scores each based on an ordinal scale from 1 to 20.

Each complete block of three tasks was counterbalanced for text setting and content type. Within each block, tasks were always presented in the following order: editing, scanning, and then reading to simplify the test design (at the cost of losing the ability to examine task order effects in detail, which was not considered central to this study). All 38 participants completed two blocks of tasks in the following manner:

Table 1. Task design

| | | | | | | |
|---|---|---|---|---|---|---|
| **Block 1** | Edit | TLX | Scan | TLX | Read | TLX |
| 2 to 5 Minute Break | | | | | | |
| **Block 2** | Edit | TLX | Scan | TLX | Read | TLX |

### *Equipment*

Participants all used the same Dell Latitude C840 laptop, with the choice of using a mouse attached peripherally or the mouse-equivalents on the laptop keyboard. The screen size of the laptop was 15 inches, with a display setting of 1600x1200 pixels. ClearType was gamma tuned for the laptop to optimize its setting for this specific laptop.

The fonts used for each of the tasks reflect the default styles typically used in home or office computer environments. The editing task was performed in Microsoft Word 2002 and used Times New Roman in 12 pt size. The scanning and reading tasks used Arial 12 pt size and were completed using Microsoft Excel 2002 and Mozilla Firebird 0.6.1, respectively. For line length, the editing task used the default 8.5 inch "Page Layout View" display width (which used about half the available horizontal screen space). The scanning task was slightly longer in length, using approximately 1000 pixels across. The reading task line length was based on the fit of text across 800 pixels (as opposed to the full 1600 pixels available) to control for the effect of lengthy lines of text on reading ability. The participants' responses were automatically recorded as they pressed the "Next" or "Done" button to proceed.

### *Procedure*

Each participant was seated in a closed-door classroom environment facing the laptop. The moderator explained the multi-part procedure for the session by showing examples of each task and conducted a walkthrough of the procedure with each participant until they were comfortable with the requirements and flow. Participants were asked to begin the session by pressing a “Start” button and proceeding through the first block of three tasks at their own pace. At the end of Block 1, the stimuli and

the moderator prompted a 2 to 5 minute break, before the participant proceeded with the second block of three tasks. All participants completed all tasks. The moderator was always present in the test room but sat out of view of the participant.

*Pilot Studies*

Two rounds of pilot studies identified key areas for improvement in the method. In the first, with 10 users, we found that the scanning task was taking approximately 100 seconds to complete and appeared to be relatively easy for participants. To increase this task's difficulty, the staff code combinations that the participant had to locate were transposed and re-located onto differing days and times.

In the second set of pilots it was noted that by Block 2 (the second set of three tasks), participants often became fatigued or bored. This symptom was shown through participants either rushing or lagging through the last tasks. To counter the effects of participant fatigue, after Block 1 (the first three tasks), the short break was added which allowed participants to move away from the computer, walk-around and chat with the moderator.

## Results

*Task Speed*

The time it took the participant to complete each task in seconds was recorded automatically. The results are summarized for each task and each condition in Table 2. The "On" condition for each task represents when ClearType was present, while "Off" signifies regular text.

The data indicate comparable speed of performance for both the editing and scanning activities, with similar variance. For the reading task, differences between ClearType and regular text is 22 seconds on average and 850 seconds in total across 38 participants, indicating that reading occurred approximately 5.1% faster with the enhanced display.

Table 2. Summary data for task completion time

| | Edit On | Edit Off | Scan On | Scan Off | Read On | Read Off |
|---|---|---|---|---|---|---|
| **Mean** | 295 | 291 | 190 | 189 | 437 | 459 |
| **StdDev** | 87 | 80 | 86 | 72 | 128 | 130 |

A repeated measures analysis of variance (ANOVA) was conducted with display type as the primary factor and task completion time (in seconds) as the dependent measure for all three tasks. As can be seen in Table 3, ClearType has no significant effect on editing and scanning tasks but led to significantly faster reading of documents [$F_{(1, 34)} = 4.48$, $p < .05$]. These results suggest ClearType offers significant speed benefits to users for tasks that involve continuous reading of lengthy text on screen.

Table 3. ANOVA results for task completion time

| Task | Sum of Squares | Df | Mean Square | F | P |
|---|---|---|---|---|---|
| EDITING | 259.57 | 1 | 259.57 | .19 | .66 |
| Error | 45456.69 | 34 | 1336.96 | | |
| SCANNING | 50.78 | 1 | 50.78 | .02 | .89 |
| Error | 88220.37 | 34 | 2594.72 | | |
| READING | 9973.10 | 1 | 9973.10 | 4.48 | .04 |
| Error | 75762.53 | 34 | 2228.31 | | |

*Accuracy Scores*

Accuracy of each task was calculated according to the criteria stated in the Experimental Design sub-section. Table 4 shows the average number of correct answers by participants in each condition across all tasks. The editing task had a possible 16 correct answers, the scanning task had 5 correct answers, and the reading had 2 correct answers possible.

A repeated measures analysis of variance (ANOVA) was conducted with display type as the primary factor and accuracy score as the dependent measure for all three tasks. As can be seen in Table 5, ClearType has no significant effect on accuracy of activities in any of the three tasks.

Table 4. Summary data for task accuracy

| | Edit On | Edit Off | Scan On | Scan Off | Read On | Read Off |
|---|---|---|---|---|---|---|
| **Mean** | 15.68 | 15.47 | 4.21 | 4.24 | 1.65 | 1.50 |
| **StdDev** | .57 | .83 | .90 | .97 | .64 | .49 |

Table 5. ANOVA results for accuracy

| Task | Sum of Squares | Df | Mean Square | F | P |
|---|---|---|---|---|---|
| EDITING | .76 | 1 | .76 | 1.64 | .21 |
| Error | 15.67 | 34 | .46 | | |
| SCANNING | .43 | 1 | .43 | .67 | .42 |
| Error | 21.83 | 34 | .64 | | |
| READING | .01 | 1 | .01 | .03 | .86 |
| Error | 10.11 | 33 | .31 | | |

For the reading task, in addition to the two multiple-choice questions, participants answered one open-ended question that required summarization of the article. As a gauge of interest and engagement in the article read, the open-ended question was coded based on the number of words the participant wrote. The mean numbers of words written for the open-ended question were 53 for ClearType-On and 61 for regular text. There were no obvious quality differences across conditions in the answers provided.

### *Mental Effort and Visual Fatigue Scores*

After each of the six tasks, participants completed the NASA-TLX subjective rating scales. The results are summarized across scales and tasks in Table 6. Note, for all scales (except 'Performance'), a higher score represents greater perceived demand or effort. A high score for Performance indicates that the participant felt that they had done well on the task.

In general, scores on the ClearType condition are marginally better e.g., ratings of perceived effort averaged 6.70 for ClearType versus 7.02 for regular text (i.e., it was easier to read with ClearType) and a combined unweighted mean for the most relevant scales of mental and temporal demand, perceived effort and frustration yields a mean score of 7.93 for ClearType and 8.32 for regular displays, all indicating minor lessening of effort for ClearType users. However, these are non-significant differences, as confirmed with one-way t-tests.

The visual fatigue scale was an added extra that we employed, since visual fatigue has been a claimed problem for many electronic text applications. Using an identical 20 point scale to the TLX, this also revealed no difference between conditions. The general perceptions of effort and demand should have been highest for reading for comprehension, and the ratings support this, though it is not possible to remove task order effects from these data.

Table 6. Summary data for TLX scores

| | **Edit On** | **Edit Off** |
|---|---|---|
| Mental Demand | 4.84 | 5.05 |
| Physical Demand | 4.11 | 4.11 |
| Temporal Demand | 5.08 | 4.84 |
| Effort | 4.87 | 5.39 |
| Performance | 16.16 | 16.45 |
| Frustration | 4.76 | 4.26 |
| Visual Fatigue | 6.05 | 6.05 |
| | | |
| | **Scan On** | **Scan Off** |
| Mental Demand | 6.13 | 6.32 |
| Physical Demand | 3.95 | 4.16 |
| Temporal Demand | 5.53 | 5.42 |
| Effort | 6.34 | 6.76 |
| Performance | 16.29 | 16.18 |
| Frustration | 6.18 | 5.71 |
| Visual Fatigue | 7.37 | 7.53 |
| | | |
| | **Read On** | **Read Off** |
| Mental Demand | 8.53 | 9.05 |
| Physical Demand | 4.76 | 5.00 |
| Temporal Demand | 8.68 | 8.82 |
| Effort | 8.89 | 8.92 |
| Performance | 14.58 | 13.97 |
| Frustration | 5.61 | 6.47 |
| Visual Fatigue | 9.11 | 9.29 |

## Discussion

The results of this experiment indicate that improving the image quality of text on screen can have a significant advantage for users performing traditional reading activities in texts of approximately 2000 words. Enhanced displays yielded a greater than 5% advantage in terms of reading speed while comprehension of content was not affected. These results suggest that for typical computer users performing this type of task, the use of ClearType technology is advantageous.

The lack of difference in the editing and spreadsheet scanning tasks is interesting. For the editing task this suggests, as hypothesized, that for rapid back and forth visual interactions between paper and screen, the advantages in image quality yielded by ClearType may not be highly influential. This may be partly explained by the break in visual processing between text and screen and the relative ease with which target information could be located on screen in this task. Since participants in this scenario were not processing the information for comprehension or even reading it in any substantive manner, only superficial viewing of the display was required. For this task, gross characteristics of letter location and shape were all that was required and the remainder of the task involved the more physiological act of moving the cursor to the insertion point and typing a correction. Observation of participants suggests that the physical components of editing (examining document, scrolling, highlighting and typing) dominated task performance and it is possible that the time taken for such actions is swamping any advantage that may exist for

display type in this task. This is an issue that warrants further investigation.

A close examination of the speed data for the editing task indicates a somewhat bimodal distribution however, with thirteen of the participants forming a sub-group which performed slower in both ClearType and regular text conditions. A post-hoc analysis of this group alone indicated that these participants were significantly faster in editing with ClearType [$F_{(1, 9)} = 7.81$, $p < .025$] which may suggest that ClearType has greater benefits for a certain class of user e.g., one with less task experience, perhaps, or some other distinguishing variable that separates these users from the majority. This can only be speculated upon with the current data, but again, is an issue for further research.

For spreadsheet scanning, where the participant was required to spend more time looking at the screen alone to search for targets, it was expected that image quality would directly affect performance, at least in terms of speed. However, in this study, no such advantage was observed. It is possible that the tasks were not taxing enough for participants although the initial pilot test led to an increase in the difficulty of the task employed and observations of participants lend support to the argument that these tasks were challenging. More likely, as with the editing tasks, the requirement that all participants type their answers may have introduced extra activity beyond the visual scanning component. The timing mechanism was set to run until the user typed the answer into the required field and hit a 'Done' button. Again, observers noted that some users spent a large proportion of their time on this task being sure to correctly enter their answers. Including the time for this may have diluted any effects for image quality here. We intend to explore this in a follow-up study.

The advantage to ClearType for the reading task is interesting when considered in terms of typical user activity. Reading for comprehension is both a perceptually and cognitively-intensive act. The basic interaction in this scenario is extended eye-on-text, serial reading aimed at following the logic of an article in order to determine its meaning. Here, the impact of image quality is maximized since it directly affects the dominant user activities throughout the interaction. Time on this task averaged more than six minutes, substantially longer than either of the other tasks, providing an extended test of image quality on performance. As hypothesized, it is for this activity that the impact of ClearType is most clearly demonstrated. We note, however, that this contradicts the findings of Tyrrell et al. (2001) who reported no significant difference in reading speed for users of ClearType in a comparison test that involved reading for up to one hour.

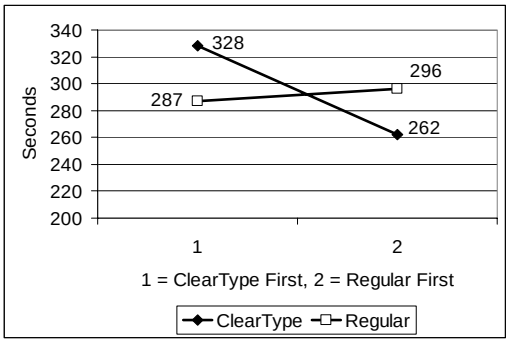
The experimental design placed substantial demands on participants. Direct observations of the participants indicated that by the end of the second block of tasks, some participants were exhibiting signs of boredom or fatigue, while others seemed to have learned from the first block and performed more confidently in block two.

As part of our post-hoc analysis we compared performance (speed) for participants on their first and second blocks. As shown in Table 7, participants generally performed faster, and variance diminished, on their second block of trials, suggesting a general performance improvement with practice on the editing [$F_{(1, 34)} = 12.64$, $p < .01$] and scanning tasks [$F_{(1, 34)} = 10.83$, $p < .01$] but not on the reading task [$F_{(1, 34)} = 1.07$, $p > .30$]. The "1" and "2" in Table 7 indicate Block 1 and Block 2.

Table 7. Summary data for task completion time

| | **Edit 1** | **Edit 2** | **Scan 1** | **Scan 2** | **Read 1** | **Read 2** |
|---|---|---|---|---|---|---|
| **Mean** | 307.87 | 279.32 | 210.95 | 168.55 | 454.34 | 441.18 |
| **StdDev** | 86.57 | 77.52 | 82.89 | 69.84 | 131.01 | 127.89 |

Examining these differences further across text condition (ClearType or Regular) adds further insight. As shown in Figures 3 and 4, the gain in performance for Block 2 occurs only for participants who experienced ClearType in that block, with regular text scores appearing flat across blocks. This suggests there may be a delayed advantage for ClearType that only emerges when users are practiced on a task.




**Fi**gure 3: Editing task times across conditions

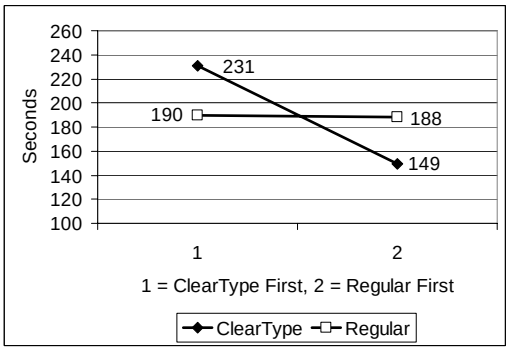



Figure 4: Scanning task times across conditions

On the contrary, for the reading task, the major differences are found in Block 1, where participants first read a text for comprehension. In fact, by Block 2, the differences between ClearType and Regular displays have disappeared for the reading task, a finding we attribute to the general tiredness of participants by this time, since all users completed this task last. We intend to follow this up in a further study.

Finally, the TLX data provide no indication of general increases in perceived effort at the end, nor is there a reported increase in visual fatigue. While this may raise some doubts about the utility of TLX in this context, a more plausible explanation may be a general disinterest in completing the scales after the first one or two tasks (participants had to complete TLX six times in all for this experiment, which our experience suggests is too much).

## Conclusions

Real world information tasks are a complex mix of physiological, perceptual, cognitive and social acts and it is certain that display characteristics will only impact parts of the tasks. ClearType seems to offer maximum benefits for activities requiring extended periods of visual processing of electronic text. Quick scanning and tasks involving multiple back and forth visual moves between media do not appear to be so affected by this screen enhancement. We plan further studies to explore this issue in more detail.


## ACKNOWLEDGMENTS

This works was supported by a research grant from Microsoft Research. The authors thank Kevin Larson for his insights into ClearType, and Kai Mantsch for technical support in the implementation of this study.


## REFERENCES


Bernard, M., Chaparro, B., Mills, M. and Halcomb, C. (2003) Comparing the effects of text size and format on the readability of computer-displayed Times New Roman and Arial text. *International Journal of Human-Computer Studies*, 59,6, 823-836.

Boyarski, D., Neuwirth, C., Forlizzi, J. & Harkness-Regli, S. (1998). A study of fonts designed for screen display. In *Proceedings of CHI '98* (pp. 87-94). New York: ACM Press.

Dillon, A. (2004). *Designing Usable Electronic Text*, 2nd edition. Boca Raton:CRC Press.

Gould, J.D., Alfaro, L., Finn, R., Haupt, B. & Minuto, A. (1987). Reading from CRT displays can be as fast as reading from paper. *Human Factors*, 29(5), 497-517.

Muter, P. and Mauretto, P. (1991). Reading and skimming from computer screens and books: The paperless office revisited?. *Behaviour & Information Technology*, 10(4),257-266.

O'Hara, K. and Sellen, A. (1997). A comparison of reading paper and on-line documents. In *Proceedings of CHI '97* (pp. 335-342). New York: ACM Press.

Tyrrell, R.A., Pasquale, T.B., Aten, T., & Francis, E.L. (2001). Empirical evaluation of user responses to reading text rendered using ClearType technologies. *Society for Information Display 2001 Digest of Technical Papers*, 1205-1207.