# Investigating font effects in continuous reading and a Lexical Decision Task

This paper presents two studies investigating font advantage between a serif (Times New Roman) and sans serif (Arial) in reading Irish. The first study involved 65 children where there were eight paragraphs in Irish taken from age appropriate class textbooks.

The paragraphs were presented in either Arial or Times New Roman (TNR) at normal contrast levels[1]. A clear advantage was found in reading material printed using the Arial. Longer gaze durations were recorded while reading of TNR. The second study investigated this advantage further in a more focused manner, it used a lexical decision task (LDT) where subjects were presented stimuli (words and non-words) randomly in either left or right parafovea. The subject's task was to make a lexical decision as quickly as possible. The contrast level of the display was set for each individual during a practice period to give approximately 50% performance accuracy. Additionally, subjects were required to provide a confidence rating between one and five, indicating how confident they were in their decision. From these responses Receiver Operator Characteristic(ROC) curves were generated along with the distribution of responses and accuracy per subjects of the responses. These provide a representation of the accuracy of each subject throughout the LDT. Results showed that LDT performance was more accurate for words presented in the right visual field and that the Arial font gave superior performance compared to TNR thus supporting the first study's findings.

## Methods and Materials

### Experiment One - Apparatus

Experimental materials were presented on a 21-inch Samsung SyncMaster 1100MB CRT monitor controlled by an nVidia Geforce FX5200, 256MB RAM graphics card with a refresh rate of 100 Hz non-interlaced. The resolution of the monitor was 1024 x 768 pixels and the screen subtended 42.8(41.4 cm) horizontally and 32.1(30.3 cm) vertically. This results in 24 pixels per deg at a viewing distance of 52.6 cm. Eye movements were recorded using an EyeLink II (SR Research Ltd.) head-mounted, video-based eye tracker. The EyeLink II has a maximum sampling rate of 500 Hz. It was calibrated for each subject by instructing them to fixate a series of nine dots (0.4 diameter) displayed in a 3x3 grid pattern on the screen. Subjects were seated without their heads stabilised by a chin-rest. This was due to the subjects being children and the chin rest hindered them in reading normally. They viewed the display binocularly, though only their right eye was calibrated. Stimuli were presented and controlled by a display PC and eye movement data were collected using a separate recorder PC.

[1] The screens background was set to RGB (255,255,255) whilst the text displayed during the experiment was presented with RGB value (0,0,0)

### Experiment One - Subjects

The experiment involved a total of 65 students from two National Schools participating in the experiment. Subjects were drawn from the 4th and 6th grade of primary school respectively. They were categorised into three levels of proficiency in the Irish language by their teacher (Low, Medium, High).

### Experiment One - Materials and Design

Materials were taken from the schools syllabus for 6th and 4th class, it consisted of short stories and poems (Dooley, 2004). Subjects were presented with material appropriate to their grade. The material was presented using the exact layout in which it had been printed. The purpose of this was to make the reading task as similar to normal reading as possible.

Eight paragraphs were presented to subjects, each contained on average 129 words. Four paragraphs were displayed using a serif(Times New Roman) font and the other four paragraphs were presented using a sans-serif(Arial)font. The contrast of the screen was set to normal luminance of RGB value (255,255,255) and the order of paragraphs was randomised to avoid bias.

Upon completion of reading task the subject used a game pad to signal completion, they were then presented with four questions relating to the paragraph that had just been read. Each question required a yes or no answer and the subject answered by pressing a specified button on the joy pad. The purpose of the questions was to provide the subjects with an impetus for carrying out the reading task in a focused manner. The subjects answers were not recorded and are not part of any results presented in this paper, eye movement data was only recorded during the reading of the paragraphs.

### Experiment One - Procedure

Prior to the start of the experiment, subjects read a short set of instructions in which they were told to read eight paragraphs. They were told that after the paragraphs, they would be presented with four questions relating to the material they had just read and their task was to indicate a yes or no answer using buttons on a joy pad. They were asked to read at their normal rate.

Each trial started with a fixation dot (0.4 diameter) that appeared on the left of the screen, adjacent to where the first character of the stimulus paragraph would be presented. Subjects initiated each trial by pressing an assigned button on the game pad. The EyeLink II system then performed a drift correction to correct for possible shifting of the head-mounted tracking system. When the drift correction was made, the fixation dot disappeared and the stimulus paragraph would appear on the screen. The size of the text was set at 20pt.

### Experiment Two - Apparatus

The apparatus used in the second experiment was identical to that dsecribed in the previous section. The only difference was that for the second experiment the subjects required a chin rest, due to the experimental procedure.

### Experiment Two - Subjects

The experiment involved a total of 11 University students. All subjects were native English speakers.

### Experiment Two - Materials and Design

Materials were taken from the British National corpus (Burnage & Dunlop, 1992). The length of the words and the frequency varied. There were three word lengths, four, six and eight along with two frequency levels, high and low.

### Experiment Two - Procedure

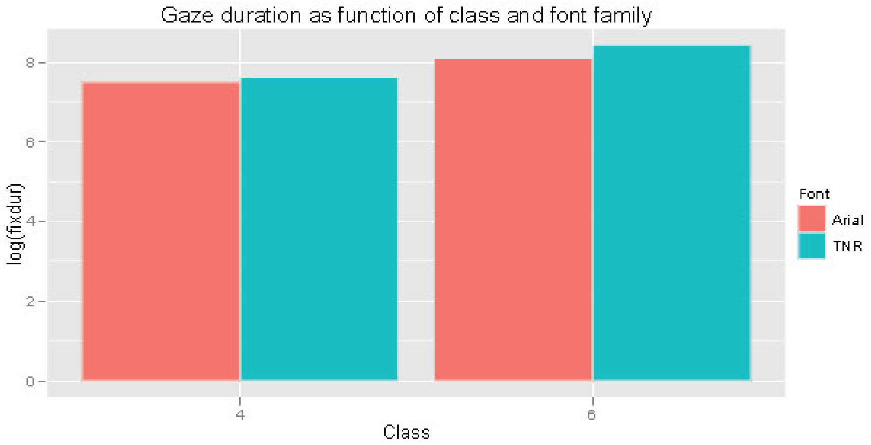
The experiment consisted of two phases, a practice phase and a trial phase. The practice phase determined a contrast level that gave 50% accuracy[2]. Once that level was determined it was then used for the experimental phase of 120 trials. The materials were presented randomly either left or right of the centre fixation point. The subjects would then make a lexical decision without fixating on the word, they then had to make the decision in the fastest time possible. Should they fixate the word, it would then be masked using the hash character.

## Results

### Experiment One - Continuous Reading Task

The main finding is that there is a clear and unambiguous advantage to reading material printed in Arial (a san serif font) compared to reading the same text in Times New Roman (TNR; a serif font) in a normal contrast environment. Crucially, while reading text in TNR, the 4th and 6th class students had longer gaze durations and made significantly more refixations of words. This is a generally accepted indicator of local reading difficulty. This is in line with previous results from (Legge et al., 1987), where reading rate for a Sans-serif font(Courier) was faster than that of reading a Serif font(Times New Roman). Moreover, the more re-fixations a reader makes, the slower their overall reading rate.

[2] The formula to determine accuracy was the following.ACC=(TP+TN)/(P+N) where TP and TN corresponds to true positives and true negatives respectfully. P and N refers to the number of positive and negative instances.

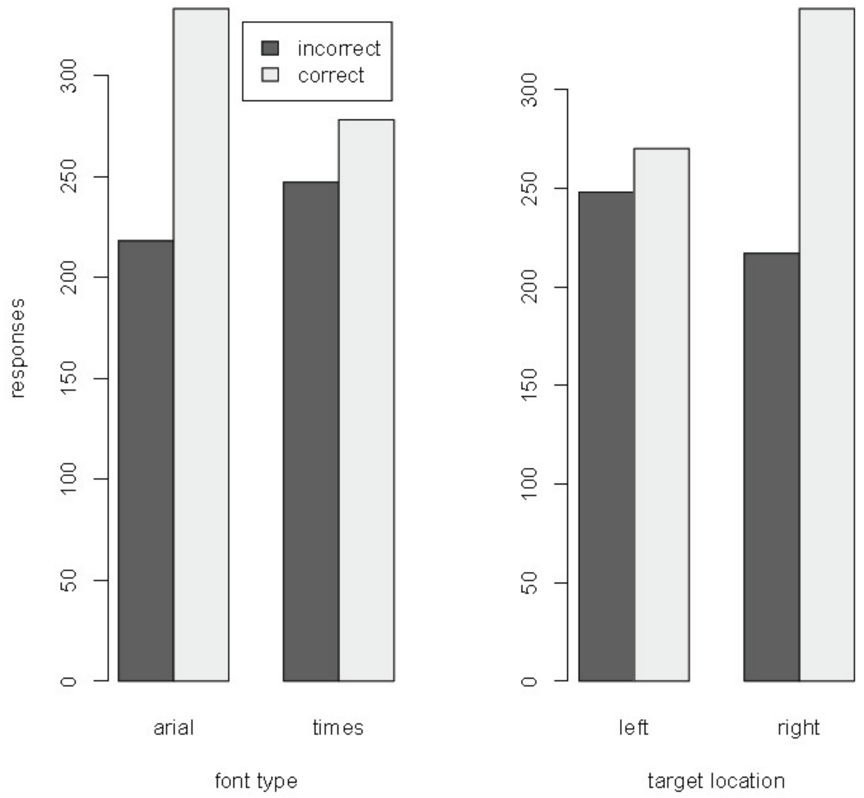


*Log gaze duration on words ranging from 2 through 11 characters as a function of class level and font family.*

The pattern of results in figure above reveals a number of interesting features. For both 4th and 6th class students, text shown in Arial is read faster. Moreover, the effect is more pronounced for 6th class students. Both the differences between classes and fonts as well as the interaction between these two factors are statistically significant (belied to some degree by the apparently small differences in figure *Log gaze duration on words ranging from 2 through 11 characters as a function of class level and font family*). A curious finding is that the 4th class pupils read the experimental texts significantly faster than the 6th class pupils. We attribute this to the fact that the older students read more carefully in order to answer correctly the questions about the text, whereas the younger students struggled with the level of Irish in the texts and tended to skim the paragraphs. The texts were chosen to be intermediate in difficulty between 4th and 6th class. The assumption of text difficulty is also supported by the poorer performance in answering questions found with the younger cohort of students. Note that we use the log of the gaze duration to correct for the skewed nature of the distribution caused by some very long durations. This is not unusual in the case of young readers.

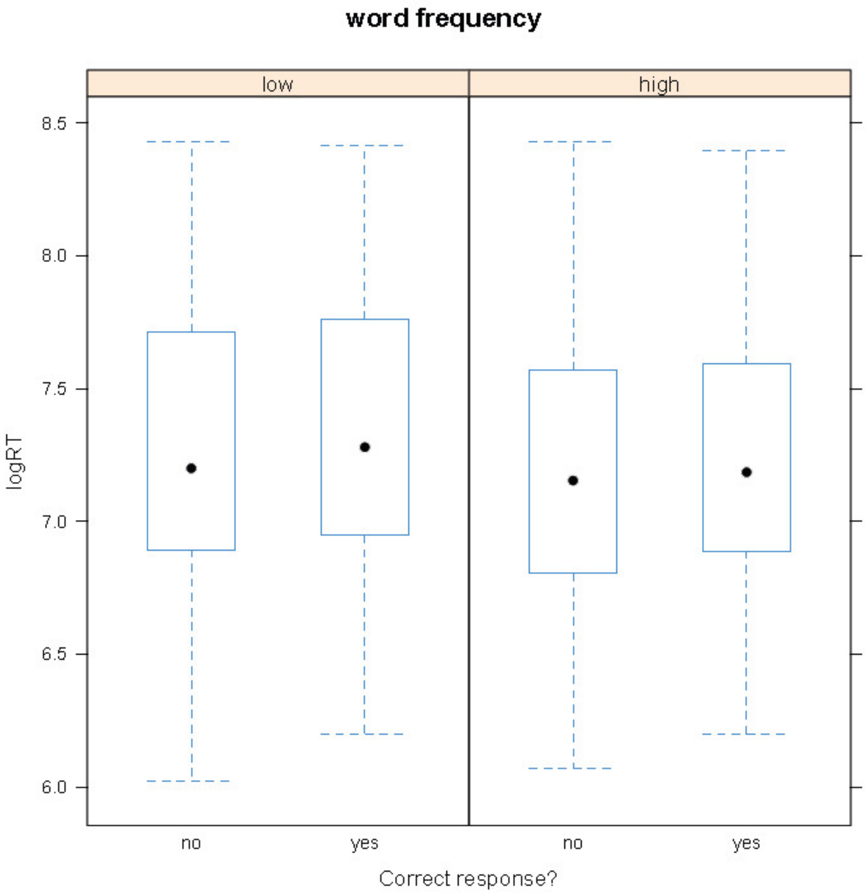
### Experiment Two - Lexical Decision Task

Data from this experiment was analysed using a linear mixed effects model (Kliegl, 2007). In figure **??** we can see correctness of response as a function of font and word location.



*Correctness of response as a function of font type*

A significantly greater amount of correct responses occur when the stimuli are presented in Sans-serif and to the right (|t |>2). The correctness of response when stimuli appear to the right is not surprising as the subjects were readers of western orthographies thus their perceptual span to their right is larger than their left Rayner et al. (2009).



*Correctness of response as a function of frequency*

Figure **??** displays decision times for both low and high frequency words with a significant elevation in decision time for words that are of low frequency. Higher response times for low frequency words is not surprising as words of higher frequency are more often skipped in reading (Rayner, 1998). The increase in decision times for low frequency words is similar to the findings of Inhoff & Rayner (1986), where subjects fixated longer on words of low frequency. However is worth pointing out that within this experiment subjects were unable to fixate words. This frequency effect is not an age related effect as subjects in experiment two were adult readers. Had the subjects been children as in experiment one, then the effect could be attributed to their age.

### ROC Analysis

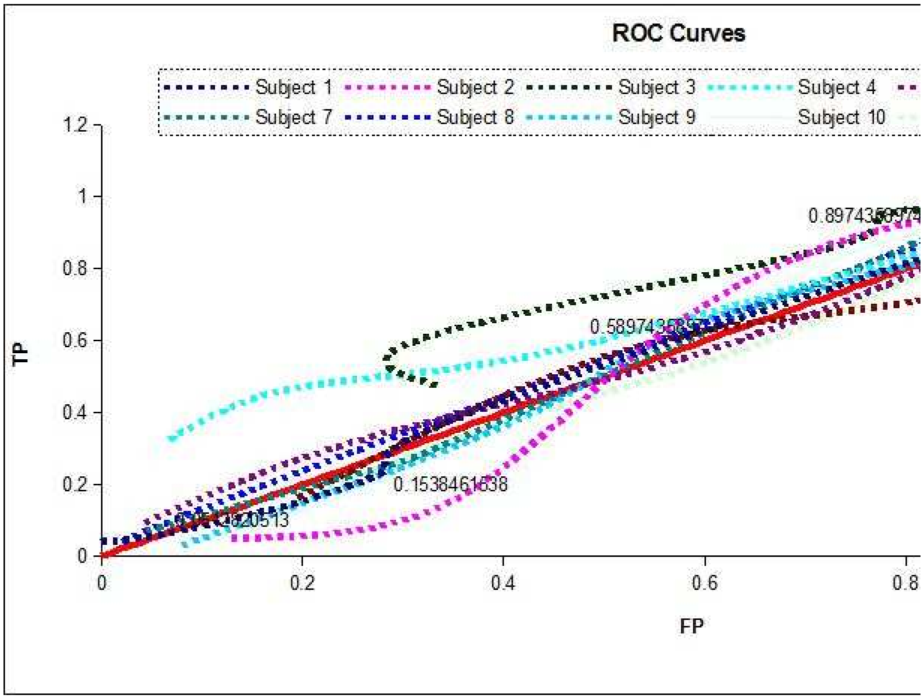
A ROC graph is a technique for visualising classifiers based on their performance. Historically ROC graphs have been used in signal detection theory(SDT) to illustrate the trade off between hit rates and false alarm rates of classifiers (Fawcett, 2006).

Since one of our experiments was a LDT, a binary classification model is used. That is, each instance $\boldsymbol{I}$ is mapped to and element of the set $\{\mathbf{p},\mathbf{n}\}$ of positive and negative class labels. (Fawcett, 2006) describes a *classification model* as a mapping from instances to predicted classes. Once a subject is presented with a stimulus there are four possible outcomes based on their classification. Since the instance could be a word or non-word, a word is described as being positive and a non-word is negative. If the instance is positive and it is classified as being positive, it is counted as a *true positive*; if it is classified as being negative, it is counted as being a *false negative.* If the instance is negative and it clasified as negative, it is counted as a *true negative*; if it classified as positive, it is counted as a *false postive.* When creating ROC curves the two metrics of interest are the **true positive rate** and the **false positive rate**. They are plotted on the $Y$ axis and the $X$ axis accordingly. Theses are defined to be

$$tp\ rate \approx \frac{\text{Positive instances correctly classified}}{\text{Total Positives}}$$

$$fp\ rate \approx \frac{\text{Negatives incorrectly classified}}{\text{Total Negative}}$$

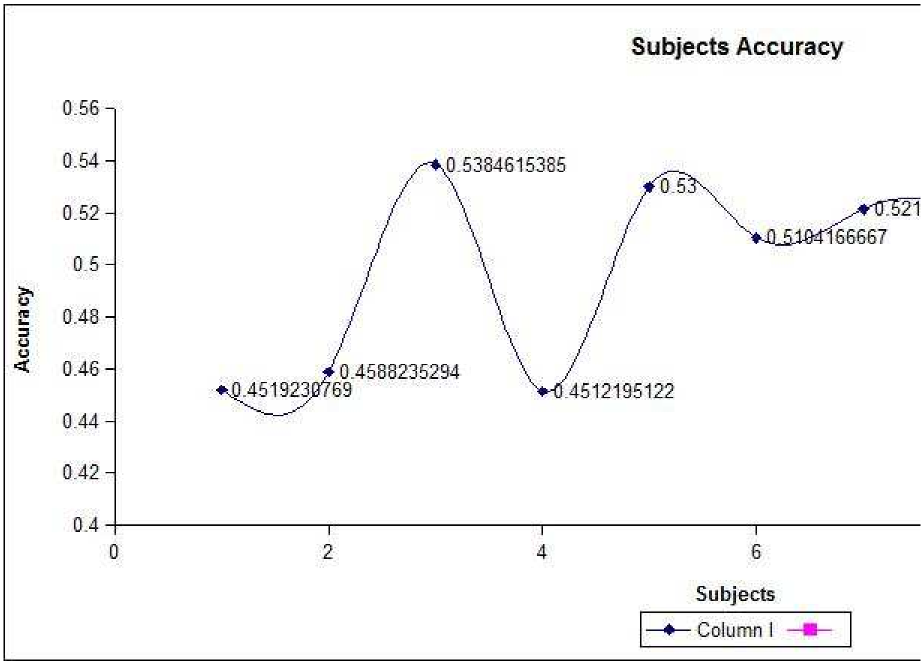
The points of the ROC are plotted each between 0 and 1 on each axis accordingly. The points of the graph are representative of the strategy used to classify the instances. Entries along the diagonal $x$=$y$ are indicative of a guessing strategy. If points are located close to the left hand side of the X-axis, they are thought to be conservative since they make few false positive classifications but make positive classifications only with strong evidence. Classifiers located in the up left are thought of as *liberal* since they make positive classifications with weak evidence (Fawcett, 2006).



*Complete ROC curves from all subjects*

In our LDT experiment we required that subjects make their decision and then rate their decision based on a confidence rating. The ROCs were plotted by summing the number of ***tp*** and ***fp*** across the ratings scales. In figure **??** we have all the generated ROCs from the subjects that took part in the LDT. Amongst them is the diagonal line where $x$=$y$ plotted in red. This line indicates random classifications. It is clear to see from the graphs that most of the subjects were not incredibly confident in their decisions.

Figure **??** depicts the accuracy of the subjects. A guessing strategy is represented by 0.5 and it is clear from the results that the subjects adopted strategies slightly more decisive than guessing. These strategies can be explained by the difficulty of the task. The subject had to make a decision in an extremely challenging environment. The contrast was adjusted to ellicit 50% accuracy and the subject was unable to fixate the stimuli presented ramdonly either side a centre cross.



*Accuracy scores of the subjects taking part in the LDT experiment*

## Discussion

The results from the two experiments are straightforward: There is a cross task and visual environment advantage when text is displayed using a sans-serif font.

Since the subjects of the continuous reading experiment were children it is reasonable to wonder whether the superior performance of the sans-serif font extends to adult readers. The answer is that it does. Not only is the advantage seen across subjects. It also extends from a normal reading environment to a much more difficult one.

The objective of setting the LDT environment to one where subjects ellicited 50% accuracy, was to determine if an advantage transended viewing environments? The results showed that it did. ROC analysis of the subjects ratings showcased the difficulty of the task in such an environment. And their accuracy values matched the desired experimental goals. Thus pointing to the strength of the effect.

Known properties of reading western orthographies were also present within the results. Response times were quicker for high freqency words, responses were more accurate to the right of the centre fixation point. These asymetries would have been expected but their existence diminish the suggestion that subjects were just guessing if one notes that accuracy scores were very close to 0.5.

But how did the serifs effect the viewing environment of the LDT compared to the previous one of the continuous reading task? Firstly, as indicated in the introduction the serifs could have acted as noise and as a result the word identification process could have been facilitated by their absence Moret-Tatay & Perea (2011). Since the presence of serifs leads to an increase in the inter letter spacing, this increase in spacing leads to greater legibility but has been shown to have no effect on reading times (Arditi & Cho, 2005). The increase in spacing in this experimental scenario however would not necessarily be seen as providing an advantage. The stimuli in the LDT were presented in the parafovea and the subject was unable to fixate them. The increase in spacing would actually push the letters of the stimuli further into the parafovea. Spacing has been noted before, (Liu & Arditi, 2001) carried out a letter recognition task where the inter letter spacing was 1.0 and 0.1 of letter height. Results showed that less error occurred when the larger spacing was used. However in this difficult viewing environment, any spacing increase or reduction would have increased the difficulty of making a lexical decision. The lack of serifs in the text that was presented in Arial may have been a key factor due to the simple fact that the visual scene is more cluttered when the text is presented in TNR. The results also undermine the argument that serifs act as guides to the readers McLean (1980).

This study suports the results of (Moret-Tatay & Perea, 2011), they also performed a lexical decision task similar to ours. Their method however was more typical of a lexical decision experiment where the stimuli were presented centrally for the subject to identify as quickly as possible. The contrast was not set to a level of 50% accuracy and the subjects were allowed to fixate the stimuli. They concluded that serifs provided no benefit to the word identification process. Our results support theirs and show that serifs do *not* facilitate lexical decisions; instead, the presence of serifs may hinder them.

## Conclusion

The results from the two experiments show a significant font advantage when materials are presented in a sans serif font. The results of the two experiments complement each other by displaying a font effect across a continuous reading task and a lexical decision task. Additionally this effect is also displayed across normal contrast and low contrast viewing environments. The results from both experiments provide additional confirmation of results found in the past by (Moret-Tatay & Perea, 2011) by virtue of the fact that both experiments in this paper were carried out in much more difficult experimental conditions than theirs (Arditi & Cho, 2005).

## References

Arditi, A., & Cho, J. (2005). Serifs and font legibility. *Vision Research*, *45*, 2926.

Burnage, G., & Dunlop, D. (1992). *Encoding the british national corpus.*

Davis, K., Woods, R., & Scharff, L. (2005). Effects of typeface and font size on legibility for children. *American Journal of Psychological Research*, 86–102.

Dooley, H. (2004). *Bualadh bos 4 & 5.* Carroll Education Limited.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*, 861–874.

Inhoff, A., & Rayner, K. (1986). Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Perception &: Psychophysics*, *40*(6), 431–439.

Kliegl, R. (2007). Linear ,mixed-effects distributed models for reading fixations. *Journal of Experimental Psychology: General, Online Supplement*, 1–22.

Legge, G., Rubin, G., & Luebker, A. (1987). Psychophysics of reading: V.the role of contrast in normal vision. *Vision Research*, *27*, 1165–1177.

Liu, L., & Arditi, A. (2001). How crowding affects letter confusion. *Optometry and Vision Science*, *78*(1), 50–55.

McConkie, G., & Rayner, K. (1975). The span of the effective stimulus during a fixation in reading. *Percetion and Psychophysics*, *17*, 578–586.

McLean, R. (1980). *The thames and hudson manual of typography.* Thames; Hudson Ltd.

Moret-Tatay, C., & Perea, M. (2011). Do serifs provide an advantage in the recognition of written words? *Journal of Cognitive Psychology*, *23*, 619.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*, 372.

Rayner, K., Castelhano, M., & Yang, J. (2009). Eye movements and the perceptual span in older and younger readers. *Psychology and Aging*, *24*(3), 755.

Rayner, K., Reichle, E., Stroud, M. J., Pollatsek, A., & Williams, C. (2006). The effect of word frequency, word predictability, and font difficulty on the eye movements of young and older readers. *Psychology and Aging*, *21*(3), 448–465.

Rubenstein, R. (1988). *Digital typography: An introduction to type and composition for computer system design.* Addison Wesley.