Information Design Journal 18(2), **118–137**

DOI: 10.1075/idj.18.2.03bei

Sofie Beier and Kevin Larson

# Design Improvements for Frequently Misrecognized Letters[1]

**Keywords: legibility, visibility, typeface design, fonts, lower-case letters, experimental study, typographic research**

**To enhance typeface legibility we studied how to improve the design of individual letters. Three different fonts were created, each containing several variations of the most frequently misrecognized letters. These variations were tested both with distance and short exposure methodologies. Creating variations within a typeface avoided confounds that occur when letters from different typefaces are compared against each other. The studies found that some variations were more legible than others despite the letters within a font having similar size, weight, and personality. The results showed that narrow letters benefit from being widened, and that x-height characters benefit from using more of the ascending and descending area. These findings can be used to improve the design of future typefaces.**

A common approach in experimental legibility studies is to compare one font against another font. A valuable critique of this method is the issue of confounding parameters between fonts, such as proportions, weight, stroke, contrast and look. With so many parameters varying in the test materials, it is difficult to identify the variables that influence the findings.

To inform the choice of typeface for signage at Heathrow Airport's Terminal 5, Robert Waller (2007) compared five different fonts: BAA Signs, Frutiger Bold, Frutiger Roman, Vialog and Stempel Garamond Italic. Waller's study found, by measuring how long it would take to recognize gradually enlarged words, that Frutiger Bold is the most legible of the five and that Vialog is less legible than either of the Frutiger variations or BAA Signs. Waller speculates that the narrow width of Vialog could be causing the font's poor performance. Unfortunately, it is very difficult to be certain why Vialog performed poorer because it is different from the other typefaces on several dimensions. BAA Signs is a serif face while Vialog is a sans serif face. BAA Signs uses a double-story g, while Vialog uses a single-story g. Both BAA Signs and Frutiger Bold are heavier in weight than Vialog. With all of these variables influencing the test material, it is not easy to identify the exact reason why Vialog was less legible than the others.

With a short-exposure method Fox, Chaparro, & Merkle (2007) investigated the performance of the letters 'e' and 'o' in 20 popular text typefaces. They found that they could measure performance differences between the 'e' in many typefaces and between the 'o' in many typefaces, but that it was difficult to make claims about why some letters performed better. With a regression analysis they showed that when the letter 'e' had a higher crossbar

it was more likely to be misrecognized. This is an excellent first step at investigating the issues that are important in identifying factors that determine legibility, but it might be because of the many differences between real-world letters that no characteristics beyond the height of the crossbar was identified for the letter 'e'.

Waller and Fox et al. both study legibility by making comparisons across a variety of different typefaces. This kind of study has the advantage of studying real world typefaces that have been optimized by type designers for a particular purpose. There are an infinite number of ways to design any particular letter, but by examining existing typefaces we see a representative sample of possible designs. The disadvantage of this kind of study is that the designs that are studied differ on many dimensions making it difficult to understand the source of the observed difference.

Our plan in this project is to take a different approach from Waller and Fox et al. Instead of studying existing typefaces that differ on many dimensions, we will create different versions of letters in the same font in order to reduce the number of characteristics that are being changed between letters, and so making it easier to understand the reasons for performance differences. However, one downside of this method is that we are not able to examine the full variety of designs that are seen in the real world; another concern of looking only at variations within a single typeface is that the conclusions may only apply to that typeface alone. However, we can be more certain that the finding will broadly apply to letter recognition by empirically testing three different fonts under the same conditions.

Like Fox et al., we focus on recognizing isolated letters because several researchers have demonstrated that part of the reading process consists of a parallel recognition of the letters in a given word (McClelland & Johnston, 1977; Rayner & Pollatsek, 1989). More recent work further suggests that out of the three mental operations: letter-by-letter, word-wholes, and sentence-context recognition, the letter-by-letter operation is the strongest (Pelli & Tillman 2007). To avoid the crowding phenomenon of interfering neighboring characters, the present study has been based on a single-letter method, where each individual character is exposed to participants and not as part of a word.

Many typographers understand the reading process and are similarly concerned with single-letter recognition. The renowned typographer Walter Tracy defined legibility as being "clarity of single characters" (Tracy, 1986, p.31). Following this designation, issues such as character differentiation, contrast, stroke angle, weight, width, resolution, and hinting, can all be influencing legibility. The variable under study in the present investigation is the differentiation of characters; the other variables stay constant within each font. A notion often emphasized by typographers is that different reading situations influence legibility in ways that are not always the same. To study the performance of letter variations, not only in relation to a specific situation but also on a more general level, the present investigation contains two test methods of threshold studies; one study is based on the short-exposure method focusing on parafoveal vision, the other focuses on recognition at a distance.

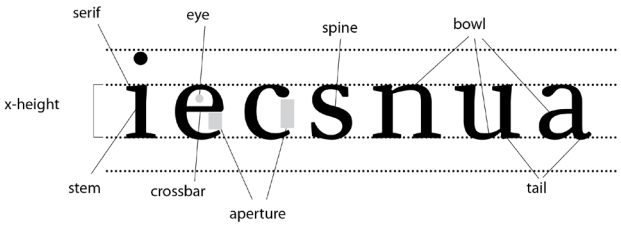



**Figure 1.** Typeface terminology.


**Figure 2.** Skeleton variations of the three fonts SpencerTest, PykeTest and OvinkTest. The fonts are all named after twentieth-century legibility researchers.

## Test Material

Studies of letter recognition tend to find similar error patterns (Geyer, 1977; Bouma, 1971; Tinker, 1964). There are two main groups of characters with high error rates. One is composed of the x-height characters of standard width, built on a mixture of straight and curved lines (e-c-a-s-n-u-o). The other group is composed of the narrow letters with a single vertical stroke and no width (i-j-l-t-f). These two letter groups are the main subjects of the present investigation.

The shapes of the skeleton variations under study are inspired by the differentiation theory put forward by Legros & Grant (1912). In a publication describing different aspects of the printing process of the day, Legros & Grant measured within a range of different fonts, the amount of overlap of similar letter pairs placed on top of each other (c-o-e, n-u, b-h, s-a, i-l). Fonts with the largest amount of overlapping areas were defined as being less legible than fonts of letter pairs with a smaller amount of overlapping areas.

### Familiar letter variations

The goal of all the letter variations is to create a greater distinction between letters. With a few exceptions, similar letter skeletons were tested on each of the three fonts. The variations of the letter 'i' in the SpencerTest and the OvinkTest faces focus on different levels of serifs. The serifs emphasize the separation of the stem from the dot, and are expected to have better legibility than the versions without serifs. Serif faces need serifs on the 'i', therefore there was no reason to test these variations in the PykeTest. For similar reasons the tailless 'u' was not tested in the PykeTest because it is aesthetically too out of place in a Serif face.

A high level of differentiation between 'n' and 'u' is expected to improve legibility. To study this hypothesis,

u2 has no tail and the bowl of versions n2 and u3 detaches closer to the middle of the stem than does versions n1 and u1. In doing so it is expected that focus will be directed towards the areas where the letters are most different from each other. A similar diagonal stroke is represented in the crossbar of version e2. This is expected to improve the recognition rate by opening up the counter.

The *Law of Closure* described by the German school of Gestalt psychologists, suggests that our perceptual system tends to complete incomplete shapes by filling out gaps. Following this hypothesis it would be expected that the smaller the aperture in 'c' and 'e' the larger the risk that the eye will close the gap and mistake these letters for 'o'. The hypothesis is further studied in open and closed apertures of the letter 's' in the OvinkTest. Following the same idea, the familiar two-storey 'a' has versions with open apertures, and versions with more closed apertures. It would be expected that closed apertures result in terminals optically joining the bowl and then lower legibility.

The one-storey 'a' was tested in the SpencerTest and the OvinkTest. Due to the dominating x-height round shape, this version would be expected to show a low level of legibility and a high number of misreadings for the lowercase 'o'. Yet, the single-storey 'a' has a skeleton that more closely resembles a handwritten 'a', and is therefore, possibly, more familiar.

**Figure 3.** Gestalt psychology's Law of Closure.

The narrow letters (l-f-t-j-i) cover a small horizontal area. It would be expected that if spread over a larger area the legibility of the characters will improve. Letters of this group all have wide and narrow versions tested.

**Unfamiliar letter variations**

Some of the tested variations were more unusual than others; these more unfamiliar versions can be divided into two main groups. One approach explores the possibility of extending the height of the character; the other the possibility of adding uppercase character shapes to the lowercase alphabet.

Many lowercase letters use neither the ascending nor descending space. The approach investigates the inclusion of the ascending and descending areas of some of the letters that do not usually make use of this space. Normally being x-height characters, the a4 and s3 move above and below this area. We know that larger sizes are more easily perceived than smaller sizes at a distance, so by extending the 'a' into the ascending area and the 's' into the descending area, it would be expected that the otherwise highly compact inner spaces of the characters open up and become more distinctive.

During the evolution of the lowercase alphabet, the early uncial pen hands mixed present-day upper and lowercase alphabets. Inspired by this tradition, the letter variations n3, e5 and t3 are uppercase letters reduced to x-height characters – the hypothesis goes that these already recognized letterforms could replace the existing lowercase versions, and still function in combination with the uppercase alphabet.

## Methods of short exposure study

The first study applied a method of short-time exposure of a single character in the parafoveal view. The findings relate to situations of reading running text.

**Participants**

There were a total of 41 participants in this study. Not all participants saw all three typefaces. 15 only saw the SpencerTest, 2 only saw the OvinkTest, 18 saw the OvinkTest and the PykeTest, 3 saw the SpencerTest and the OvinkTest, and 3 saw the SpencerTest and the PykeTest. The SpencerTest and the PykeTest were each exposed to 21 participants, where the OvinkTest was exposed to 23 participants. Most of the participants were compensated with a gratuity of Microsoft software or hardware. Some early participants received no compensation.

The participants included 26 students with art and design backgrounds from the Royal College of Art, London, and 15 students from Imperial College, London. Their ages ranged from 19 to 34, with an average age of 25.7 years. The participants came from a variety of backgrounds (British, French, Brazilian, Danish, Canadian, Swedish, Norwegian, Spanish, Slovenian, Polish), and all self-reported either normal or corrected-to-normal visual acuity. Because the mean number of errors made by participants from the two schools was not reliably different, the data from the two groups will be reported combined.

**Material**

The test material was created in Macromedia Flash MX and shown on a 15-inch MacBook Pro laptop with a screen resolution of 1440 x 900 pixels set to maximum brightness. The three fonts (SpencerTest, OvinkTest and PykeTest) were all presented with anti-aliasing at the vertical size of 45 pixels (an Em-square of about 1 cm). Since this is not a study of comparison between fonts, the three faces are not adjusted according to x-height. To minimize eyestrain caused by the background light of the screen, the background color was a shaded white (#E6E6DD) with the presented letters in black (#000000). The ambient room light was typical for an office environment.

**Procedures**

Both the foveal and the parafoveal areas are important for continuous reading (Rayner, 1978; Rayner, McConkie & Ehrlich, 1978), yet the short-exposure method applied in the present study did not detect any errors of identification when test material was placed in the foveal, and so the focus was on the parafoveal alone. Test materials were therefore located 2 cm to the right of the fixation point where participants placed their focus. Their eyes were placed at a distance of 50 cm from the screen. Each character variation within a typeface was presented 3 times per participant. To maintain an approximately equal appearance between the 26 letters of the alphabet, the 15 characters of the English alphabet that were not under investigation were each exposed 5 times – all occurring in the same random order.

The instruction was to focus on a red dot on the screen and then press the space key to trigger an exposure of a single character, which participants were asked to name. Each letter was exposed for a period of about 43 milliseconds. To ease the participant into the test, a selection of the characters not under investigation were presented as the first 5 exposures. A mask (exposed for 43 milliseconds) of randomly placed black dots followed directly after each letter exposure: this removes the afterimage on the retina and controls the timeframe in which the image in reality would appear on the retina.

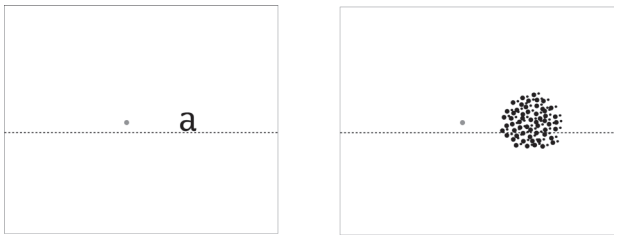



**Figure 4.** The test character (left) and the after image (right), both with a visible baseline and a dot to focus on.

Participants were informed that they would be presented uppercase and lowercase letters; they were not asked to hurry their response, as their responses were not timed.

## Methods of distance study

The second study used a distance threshold method to study the legibility of the same font letter variants. The findings relate to typefaces presented on signs viewed at a distance.

### Participants

There were 41 participants in this study, though 7 were disqualified because they did not meet the minimum visual acuity requirement of being able to recognize stimuli at a distance of 4.5 meters. This left 34 participants. All three fonts were each exposed to 20 participants: 6 participants saw the SpencerTest and the Pyketest, 6 participants saw the TinketTest and the OvinkTest, 14 participants saw the OvinkTest and the Pyketest, and 8 only saw the SpencerTest.

The participants were compensated with a gratuity of Microsoft software or hardware.

### Material

The fonts, computer, and environment were identical in the two studies.

### Procedures

In this investigation, the laptop was placed on a podium at the eye-level height of a standing person of about 175 cm. The angle of the screen was adjusted to fit the given height for each person.

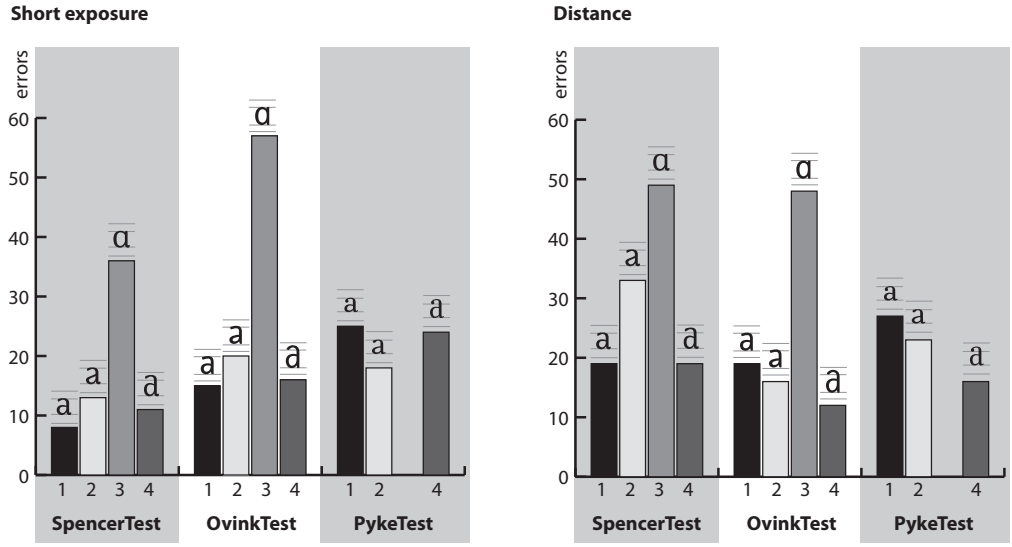
The first presented character was the letter 'd'. As identified by Tinker (1964) this character is one of the most easily recognized letters. The purpose of this first exposure was to locate the individual vision threshold. The participant was placed at a distance of 10 meters from the screen, and asked to move slowly forward until the presented letter was at the threshold of being identifiable; this was the distance – varying from 4.5-9 meters (with an average of 6 meters) from the screen – at which the individual participant was tested. From this distance, participants were asked to name each of the letter stimuli. A new letter was presented on screen after each participant response. Participants were not asked to hurry, and were permitted to take as many breaks as they felt necessary.

This method is different from the one applied in other recent distance studies, such as those by Sheedy and colleagues (2005) and the studies of the Clearview typefaces (Garvey, Pietrucha & Meeker, 1997), where the maximum distance is measured for each letter, and the distance itself becomes the data rather than the accuracy from a particular distance. However, the Sheedy and Clearview method does not identify which other letters the character tested is most likely to be misread for – a parameter that is measurable with the present method.

## Results & Discussion

In the exposure study, each letter variation for the SpencerTest and the PykeTest was presented a total of 63 times, and for the OvinkTest 69 times. In the distance study, each letter variation was presented a total of 60 times. If the participant correctly identified the presented letter, the trial was counted as correct. The inferential statistics of a chi-square distribution were conducted on the raw totals of correct and incorrect observations. Tests were only conducted between variants within a font, as it was not a goal of this investigation to compare differences between the fonts.

**Figure 5.** Letter 'a'



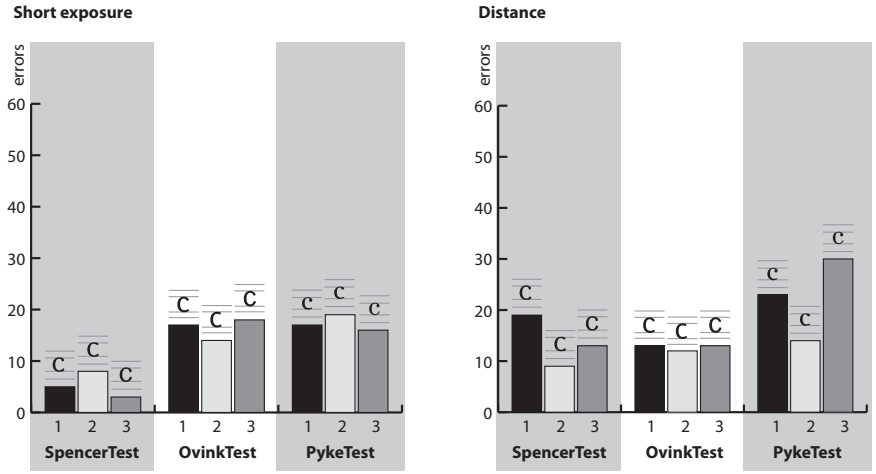

| Font | SpencerTest | OvinkTest | PykeTest | SpencerTest | OvinkTest | PykeTest |
|---|---|---|---|---|---|---|
| Study | Short Exposure | Short Exposure | Short Exposure | Distance | Distance | Distance |
| a1 errors | 8 | 15 | 25 | 19 | 19 | 27 |
| a2 errors | 13 | 20 | 18 | 33 *** | 16 | 23 |
| a3 errors | 36 * | 57 * † | - | 49 ** †† | 48 ** †† | - |
| a4 errors | 11 | 16 | 24 | 19 | 12 | 16 |
| Chi-square | $\chi^2(3)=39.8$, $p=.0001$ | $\chi^2(3)=73.87$, $p=.0001$ | $\chi^2(2)=1.99$, $p>.05$ | $\chi^2(3)=40.80$, $p=.0001$ | $\chi^2(3)=56.36$, $p=.0001$ | $\chi^2(2)=4.45$, $p>.05$ |
| Statistically Reliable | yes | Yes | no | Yes | yes | no |

* Post-hoc tests showed reliably more errors than each of the other versions.
** Post-hoc tests showed reliably more errors than each of the other versions.
*** Post-hoc tests showed reliably more errors than versions 1 and 4.
† A high frequency of misreadings for the letter 'q' (20).
†† A high frequency of misreadings for the letter 'o'. SpencerTest a3 (23), OvinkTest a3 (22).

The performance of the one-storey a3 was generally bad, with recurrent misreadings for letters 'q' and 'o'. Does this finding suggest that a one-storey 'a' should never be used? In relation to the inexperienced reader it does appear to have a purpose. Recognition is a dominant factor when learning to read; the fact that the one-storey 'a' references to the letter shape that most children learn to write, has a positive influence on the inexperienced reader (Sassoon 2001). The present study, however, focuses on the experienced reader, where references to one's own handwriting are less essential.

The hypothesis that the open aperture of a2 would improve legibility was not confirmed, showing no reliable difference in performance between a1 and a2, except for the SpencerTest distance study, where a2 performed reliably poorer than versions a1 and a4. This is an unexpected difference. A possible reason for this might originate in the shape of the bowl. The upper part of the bowl in version a2 is more diagonal in the SpencerTest than in the two other fonts; furthermore it bends slightly inwards, disrupting the dynamic movement of the curve, and making it look more like a spine than a bowl.

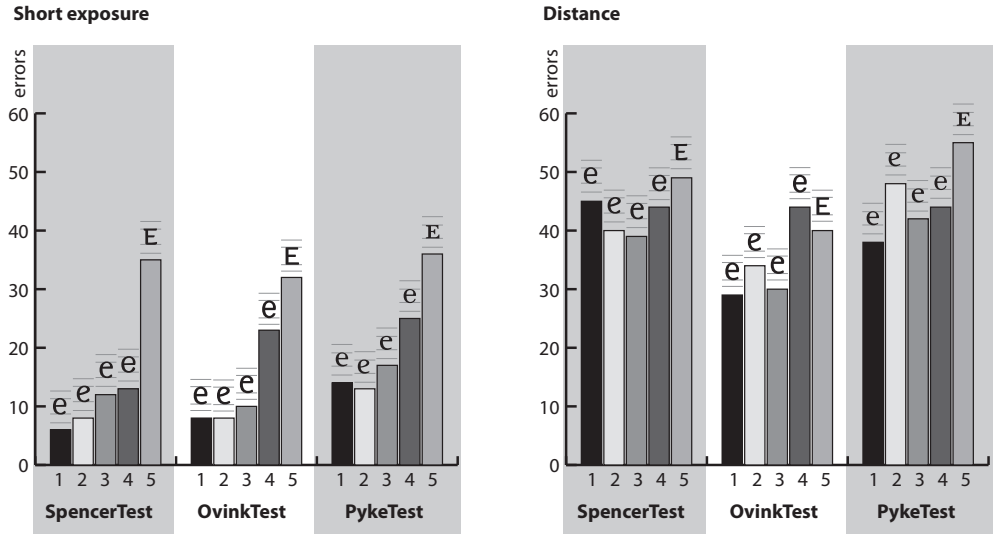
**Figure 6.** Letter 'c'




| Font | SpencerTest | OvinkTest | PykeTest | SpencerTest | OvinkTest | PykeTest |
|---|---|---|---|---|---|---|
| Study | Short Exposure | Short Exposure | Short Exposure | Distance | Distance | Distance |
| c1 errors | 5 | 17 | 17 | 19 | 13 | 23 * |
| c2 errors | 8 | 14 | 19 | 9 | 12 | 14 |
| c3 errors | 3 | 18 | 16 | 13 | 13 | 30 * |
| Chi-square | $\chi^2(2)=3.04$, $p>.05$ | $\chi^2(2)=1.09$, $p>.05$ | $\chi^2(2)=6.59$, $p>.05$ | $X^2(2)=5.82$ $p>.05$ | $\chi^2(2)=5.87$, $p>.05$ | $\chi^2(2)=39.10$, $p=.0001$ |
| Statistically Reliable | no | no | no | No | no | yes |

* Post-hoc tests showed reliably more errors than versions 2.

**Figure 7.** Letter 'e'



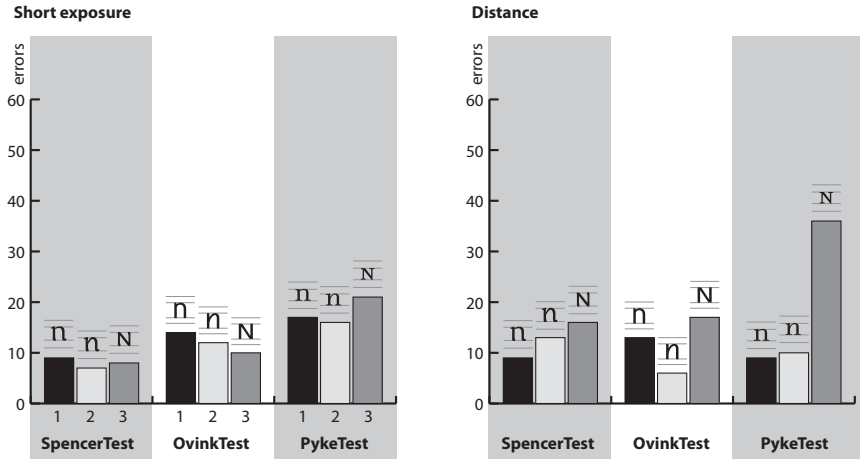

| Font | SpencerTest | OvinkTest | PykeTest | SpencerTest | OvinkTest | PykeTest |
|---|---|---|---|---|---|---|
| Study | Short Exposure | Short Exposure | Short Exposure | Distance | Distance | Distance |
| e1 errors | 6 | 8 | 14 | 45 | 29 | 38 |
| e2 errors | 8 | 8 | 13 | 40 † | 34 | 48 † |
| e3 errors | 12 | 10 | 17 | 39 | 30 | 42 † |
| e4 errors | 13 | 23 ** | 25 *** | 44 | 44 ***** †† | 44 |
| e5 errors | 35 * † | 32 ** | 36 ** | 49 | 40 † | 55 **** |
| Chi-square | $\chi^2(4)=47.94$, $p=.0001$ | $\chi^2(4)=37.82$, $p=.0001$ | $\chi^2(4)=26.43$, $p=.0001$ | $\chi^2(4)=5.43$, $p>.05$ | $\chi^2(4)=11.52$, $p=.02$ | $\chi^2(4)=15.13$, $p=.004$ |
| Statistically Reliable | Yes | Yes | yes | No | yes | yes |

* Post-hoc tests showed reliably more errors than each of the other versions.
** Post-hoc tests showed reliably more errors than versions 1, 2, 3.
*** Post-hoc tests showed reliably more errors than version 2.
**** Post-hoc tests showed reliably more errors than versions 1, 3 and 4.
***** Post-hoc tests showed reliably more errors than versions 1 and 3.
† A high frequency of misreadings for the letter 'c'. SpencerTest Short Exposure e5 (21), SpencerTest Distance e2 (21), OvinkTest Distance e5 (23), PykeTest Distance e2 (20), Pyketest Distance e3 (21).
†† A high frequency of misreadings for the letter 'o' (20).

The initial hypothesis that closed apertures in 'c' and 'e' would lower legibility was, in most tests, not confirmed in the case of the letter 'c' – only showing a statistically reliable difference between the open c2 and the more closed versions c1 and c3 in the distance study of the PykeTest. The PykeTest c1 and c3 are the only versions tested with a teardrop on top; this finding suggests that teardrops do not improve legibility at distance. It further appears that in the parafoveal vision, when the letter 'c' is viewed in isolation, the viewer registers the cut-off area in the circle regardless of the size of the area and therefore, in contrast to all existing recommendations by most typographers, showed no difference in the characters having closed or open apertures.

The hypothesis that closed apertures of e4 lower legibility was only confirmed in the OvinkTest and the PykeTest at short exposure and in the OvinkTest at distance. The three remaining familiar 'e' variations

**Figure 8.** Letter 'n'



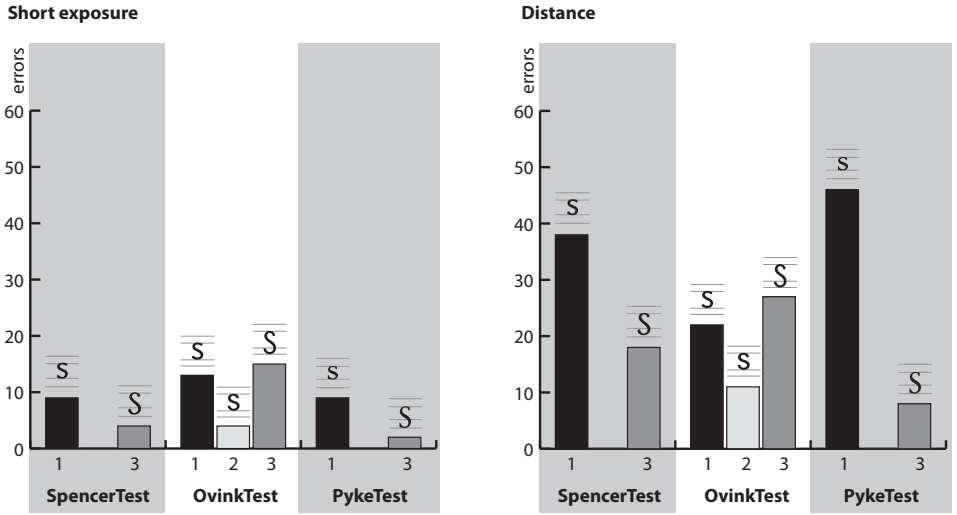

| Font | SpencerTest | OvinkTest | PykeTest | SpencerTest | OvinkTest | PykeTest |
|---|---|---|---|---|---|---|
| Study | Short Exposure | Short Exposure | Short Exposure | Distance | Distance | Distance |
| n1 errors | 9 | 14 | 17 | 9 | 13 | 9 |
| n2 errors | 7 | 12 | 16 | 13 | 6 | 10 |
| n3 errors | 8 | 10 | 21 | 16 | 17 * | 36 * |
| Chi-square | $\chi^2(2)=0.29$, $p>.05$ | $\chi^2(2)=0.81$, $p>.05$ | $\chi^2(2)=1.09$, $p>.05$ | $\chi^2(2)=2.47$, $p>.05$ | $\chi^2(2)=6.46$, $p=.04$ | $\chi^2(2)=36.81$, $p=.0001$ |
| Statistically Reliable | No | no | no | No | yes | yes |

* Post-hoc tests showed reliably more errors than versions 2.

showed no internal differences. Yet all versions except e1 demonstrate a high number of misreadings for the letters 'c' and 'o'. The e5 variation performed poorly. It appears that the upper and lower crossbars are overdominating the middle crossbar, which in some cases resulted in a high number of misreadings for the letter 'c'.

The idea that detaching the bowl from the stem would enhance legibility of versions n2 has not been

**Figure 9.** Letter 's'



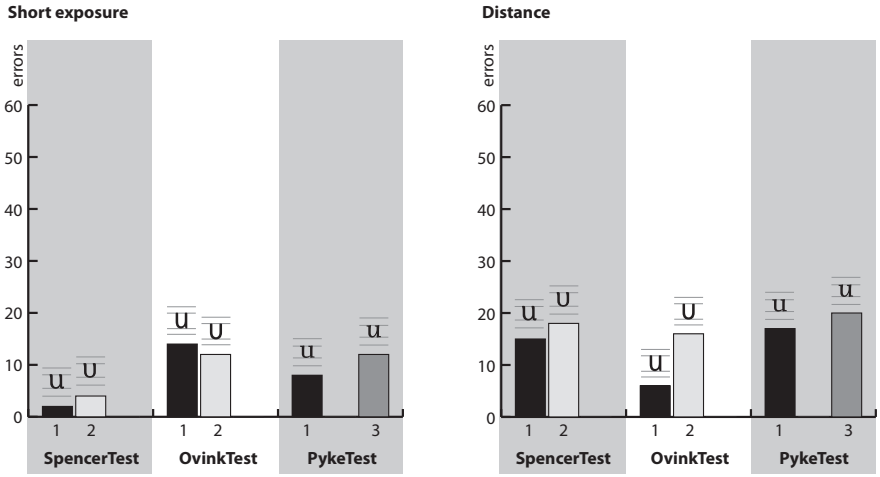

| Font | SpencerTest | OvinkTest | PykeTest | SpencerTest | OvinkTest | PykeTest |
|---|---|---|---|---|---|---|
| Study | Short Exposure | Short Exposure | Short Exposure | Distance | Distance | Distance |
| s1 errors | 9 | 13 * | 9 ** | 38 ** | 22 | 46 ** |
| s2 errors | - | 4 | - | - | 11 | - |
| s3 errors | 4 | 15 * | 2 | 18 | 27 * | 8 |
| Chi-square | $\chi^2(1)=1.37$, $p>.05$ | $\chi^2(2)=7.61$, $p=.02$ | $\chi^2(1)=5.21$, $p=.02$ | $\chi^2(1)=12.09$, $p=.0005$ | $\chi^2(1)=10.05$, $p=.007$ | $\chi^2(1)=51.39$, $p=.0001$ |
| Statistically Reliable | no | Yes | yes | yes | yes | yes |

* Post-hoc tests showed reliably more errors than version 2.
** Post-hoc tests showed reliably more errors than version 3.

confirmed – showing no statistically reliable difference over n1 in any situation. Version n3 showed no noticeable difference in most of the studies, except for the PykeTest at distance presenting a statistically reliably bad performance compared to both versions n1 and n2, and in the OvinkTest at distance favoring version n2.

The hypothesis that the closed apertures of the OvinkTest s2 would lower legibility was not confirmed. The angle of the spine seems to have influenced the performance, showing an advantage in favor of the closed apertures of s2 compared to s1 in both the short- exposure and the distance studies. The fact that the OvinkTest s1 has a diagonal spine and s2 a rounded spine might be the reason for the advantage towards version s2. It appears that the shape of the spine actually had a larger influence on the legibility of the 's' than the apertures being opened or

**Figure 10.** Letter 'u'

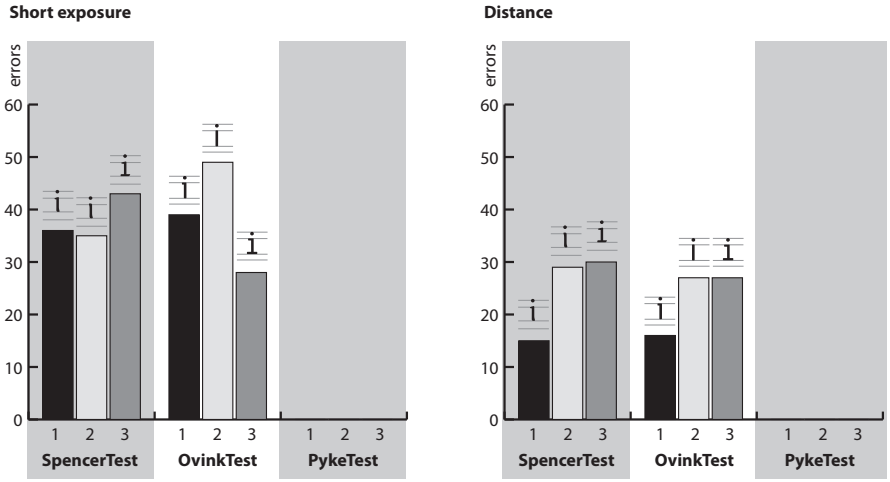



| Font | SpencerTest | OvinkTest | PykeTest | SpencerTest | OvinkTest | PykeTest |
|---|---|---|---|---|---|---|
| Study | Short Exposure | Short Exposure | Short Exposure | Distance | Distance | Distance |
| u1 errors | 2 | 14 | 8 | 15 | 6 | 17 |
| u2 errors | 4 | 12 | - | 18 | 16 | - |
| u3 errors | - | - | 12 | - | - | 20 |
| Chi-square | $\chi^2(1)=0.18$, $p>.05$ | $\chi^2(1)=0.05$, $p>.05$ | $X^2(1)=0.53$, $p>.05$ | $\chi^2(1)=0.17$, $p>.05$ | $\chi^2(1)=4.51$, $p=.03$ | $\chi^2(1)=0.16$, $p>.05$ |
| Statistically Reliable | No | no | No | no | yes | no |

closed, a finding that contradicts the recommendations of the scholar G.W. Ovink (1938), who suggested a diagonal spine of the 's'. The surprising performances of the SpencerTest a2 and theOvinkTest s2 might therefore be related. It seems that a diagonal stroke in the bowl and spine of these letters lowers their legibility and that these areas would benefit from being more rounded in shape.

The hypothesis that the legibility of 'u' would improve by differentiating the letter from 'n' was not confirmed. The tailless version u2 gave a poor performance in the OvinkTest at distance; however, in other situations it presented no statistically reliable difference from u1. Furthermore, the PykeTest version u3 showed no difference from u1.

**Figure 11.** Letter 'i'

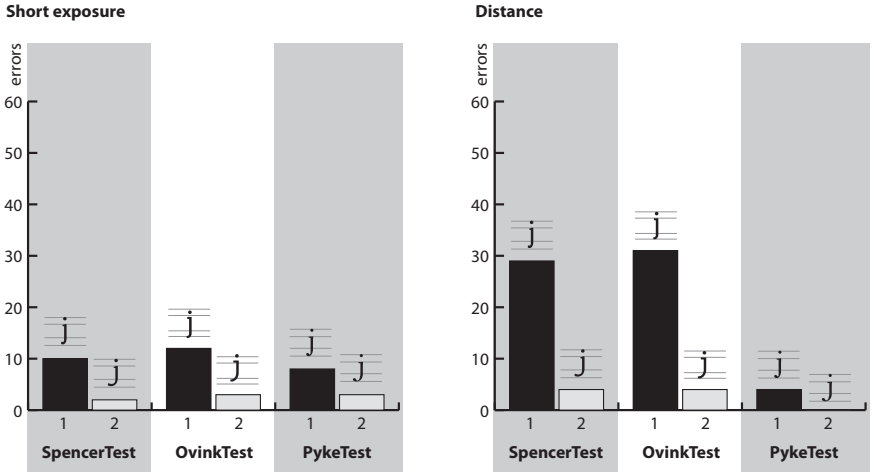



| Font | SpencerTest | OvinkTest | PykeTest | SpencerTest | OvinkTest | PykeTest |
|---|---|---|---|---|---|---|
| Study | Short Exposure | Short Exposure | Short Exposure | Distance | Distance | Distance |
| i1 errors | 36 | 39 | - | 15 | 16 | - |
| i2 errors | 35 | 49 * | - | 29 ** | 27 | - |
| i3 errors | 43 | 28 | - | 30 ** | 27 | - |
| Chi-square | $\chi^2(2)=2.52$, $p>.05$ | $\chi^2(2)=12.98$, $p=.002$ | - | $\chi^2(2)=9.68$, $p=.008$ | $\chi^2(2)=5.66$, $p>.05$ | - |
| Statistically Reliable | No | yes | - | yes | no | - |

* Post-hoc tests showed reliably more errors than version 3.
** Post-hoc tests showed reliably more errors than version 1.

The hypothesis that serifs on the letter 'i' improve legibility was confirmed for distance viewing. In both the the SpencerTest and the OvinkTest distance study, i1 with the slab serif on top was recognized more often than i2 and i3; however, only with the SpencerTest showing a statistically reliable difference, it seems as if the slab serif on top of the stem helps to clarify the letterforms, although when placed at the bottom, the character becomes difficult to identify. It appears, however, that this only happens at distance viewing, and not in the parafoveal view of short exposure, where i3 in the OvinkTest performed reliably better than i2.

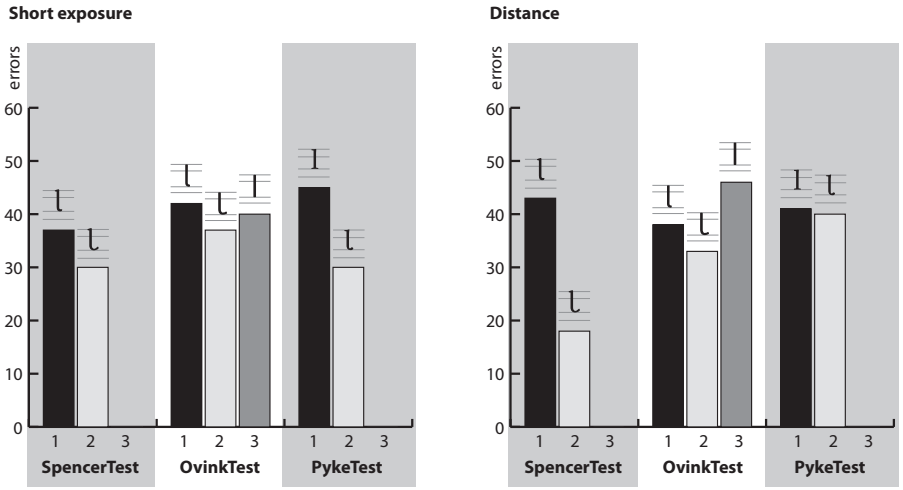
**Figure 12.** Letter 'j'




| Font | SpencerTest | OvinkTest | PykeTest | SpencerTest | OvinkTest | PykeTest |
|---|---|---|---|---|---|---|
| Study | Short Exposure | Short Exposure | Short Exposure | Distance | Distance | Distance |
| j1 errors | 10 | 12 | 8 | 29 | 31 | 4 |
| j2 errors | 2 | 3 | 3 | 4 | 4 | 0 |
| Chi-square | $\chi^2(1)=4.51$, $p=.03$ | $\chi^2(1)=4.79$, $p=.03$ | $\chi^2(1)=1.59$, $p>.05$ | $\chi^2(1)=24.08$, $p=.0001$ | $\chi^2(1)=27.27$, $p=.0001$ | $\chi^2(1)=2.33$, $p>.05$ |
| Statistically Reliable | yes | yes | no | yes | yes | no |

The hypothesis that broad characters improve legibility was confirmed overall for the letter 'j', where the broad j2 delivered a good performance on all accounts in the SpencerTest and the OvinkTest; however, no statistically reliable difference was demonstrated between j1 and j2 in the PykeTest. The broad 'j' is particularly successful because it does not introduce any new confusions. This differs from the broad form of the letter 't' which introduces a new confusion with the letter 'c'.

**Figure 13.** Letter 'l'

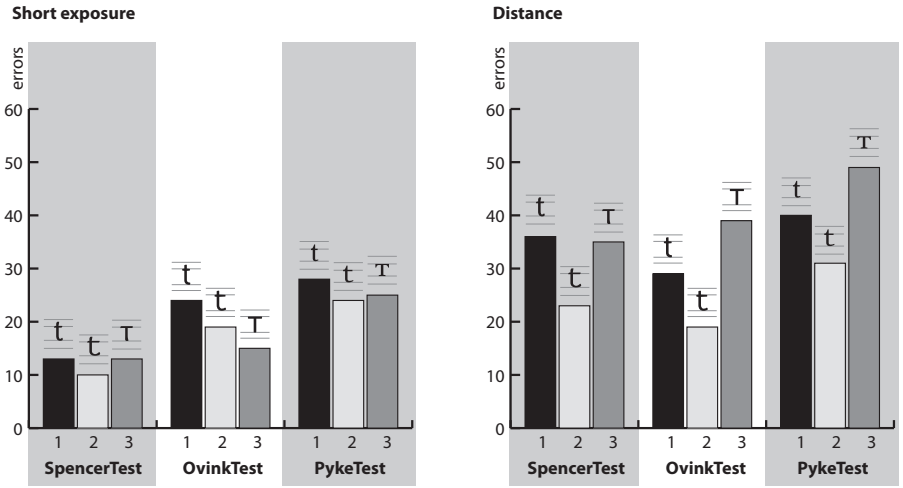



| Font | SpencerTest | OvinkTest | PykeTest | SpencerTest | OvinkTest | PykeTest |
|---|---|---|---|---|---|---|
| Study | Short Exposure | Short Exposure | Short Exposure | Distance | Distance | Distance |
| l1 errors | 37 † | 42 | 45 | 43 †† | 38 †† | 41 |
| l2 errors | 30 | 37 | 30 | 18 | 33 | 40 † |
| l3 errors | - | 40 | - | - | 46 * | - |
| Chi-square | $\chi^2(1)=1.15$, $p>.05$ | $\chi^2(2)=0.75$, $p>.05$ | $\chi^2(1)=6.46$, $p=.01$ | $\chi^2(1)=19.21$, $p=.0001$ | $\chi^2(2)=6.30$, $p=.04$ | $\chi^2(1)=0.00$, $p>.05$ |
| Statistically Reliable | No | No | yes | yes | yes | no |

* Post-hoc tests showed reliably more errors than versions 2.
† A high frequency of misreadings for the letter 't', SpencerTest Short Exposure l1 (25), PykeTest Distance (20).
†† A high frequency of misreadings for the letter 'i'. SpencerTest l1 (30), OvinkTest l1 (24), OvinkTest l3 (31).

The hypothesis that broad characters improve legibility was confirmed overall for the letter 'l'. The broad version l2 showed a reliably better performance in the Tinker distance study compared to the narrower l1, in the OvinkTest distance study compared to the straight stem l3, and in the PykeTest short exposure study compared to the serifed l1.

**Figure 14.** Letter 't'



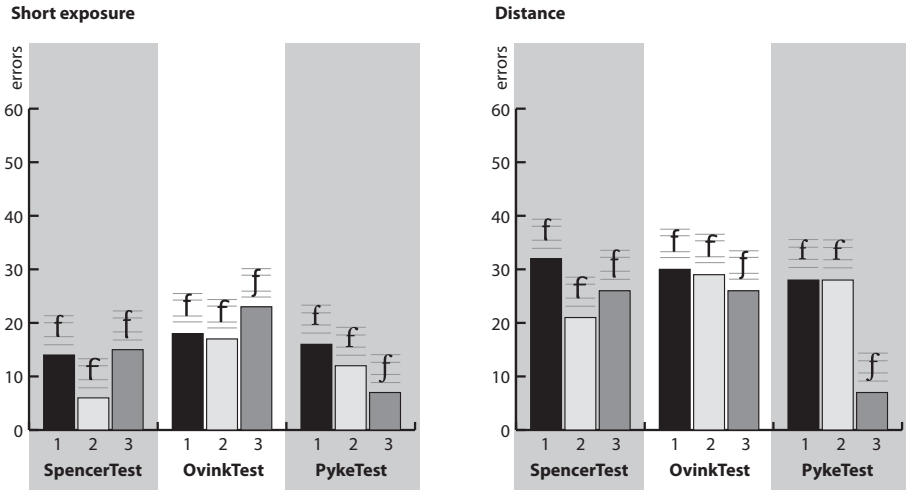

| Font | SpencerTest | OvinkTest | PykeTest | SpencerTest | OvinkTest | PykeTest |
|---|---|---|---|---|---|---|
| Study | Short Exposure | Short Exposure | Short Exposure | Distance | Distance | Distance |
| t1 errors | 13 | 24 | 28 | 36 * | 29 | 40 |
| t2 errors | 10 | 19 | 24 | 23 | 19 | 31 |
| t3 errors | 13 | 15 | 25 | 35 * † | 39 * † | 49 † |
| Chi-square | $\chi^2(2)=62$, $p>.05$ | $\chi^2(2)=2.92$, $p>.05$ | $\chi^2(2)=0.57$, $p>.05$ | $\chi^2(2)=6.99$, $p=.03$ | $\chi^2(2)=13.35$, $p=.001$ | $\chi^2(2)=12.50$, $p=.002$ |
| Statistically Reliable | No | no | no | yes | Yes | yes |

* Post-hoc tests showed reliably more errors than versions 2.
† A high frequency of misreadings for the letter 'r'. SpencerTest t3 (26), OvinkTest t3 (27). PykeTest t3 (30).

The hypothesis that broad characters improve legibility was confirmed in the letter 't' at distance viewing. All distance studies presented a small difference between t1 and t2 in favor of the latter, although only with a statistically reliable difference in the SpencerTest.

In all three fonts, the version t3 was frequently misread for the letter 'r' in the distance threshold study, and delivered a statistically reliably poorer performance compared to other versions of the 't'. On the other hand, in the short-exposure study this kind of misreading was non-existent, and the three versions of the letter performed in general quite similarly.

**Figure 15.** Letter 'f'




| Font | SpencerTest | OvinkTest | PykeTest | SpencerTest | OvinkTest | PykeTest |
|---|---|---|---|---|---|---|
| Study | Short Exposure | Short Exposure | Short Exposure | Distance | Distance | Distance |
| f1 errors | 14 | 18 | 16 | 32 | 30 | 28 * |
| f2 errors | 6 | 17 | 12 | 21 | 29 | 28 * |
| f3 errors | 15 | 23 | 7 | 26 | 26 | 7 |
| Chi-square | $\chi^2(2)=5.12$, $p>.05$ | $\chi^2(2)=1.49$, $p>.05$ | $\chi^2(2)=4.28$, $p>.05$ | $\chi^2(2)=4.11$, $p>.05$ | $\chi^2(2)=0.58$, $p>.05$ | $\chi^2(2)=21.54$, $p=.0001$ |
| Statistically Reliable | No | no | no | no | no | yes |

* Post-hoc tests showed reliably more errors than version 3.

The hypothesis that broad characters improve legibility was not confirmed in the letter 'f'. The descending f3 showed no difference in the SpencerTest and the OvinkTest; however, a reliably better performance was demonstrated in the PykeTest at distance. This result may be due to the f3 version of the PykeTest being broader in shape than the f3 version of the two other fonts. Contrary to the broad versions of the 'j' and 'l' groups, the wide f2 did not perform reliably better than any of the other tested variations.

## Conclusion

Our technique of comparing letter variations within a typeface has provided insights about letter legibility that was not previously available. Earlier studies such as those presented by Waller (2007) and Fox et.al. (2007) that examine legibility by comparing typefaces, struggle to make comparisons because every letter differs on several dimensions. Our studies are complimentary to this work by investigating letters that come from the same font with many fewer differences. This allows us to be more confident in understanding why one letter performs better than another.

Based on the findings we recommend wide versions of narrow letters. The wide j2, l2, and t2 all showed versions that performed better than their more narrow forms. The SpencerTest wide f2 also performed better than the narrow 'f'. Only for the letter 'i' was there no clear benefit for a wide form. Yet applying the broad variations of j2 and l2 in a typeface will possibly result in spacing problems: j2 will overlap with descending characters to the left, an issue causing potential trouble in the Scandinavian languages which have a high number of gj letter combinations. Version l2 would create a disrupting area of extra white space when placed to the left of another stem. When implementing these variations in a final typeface, it could be necessary to apply a number of extra ligatures and kerning pairs.

Based on the findings we can also recommend extending letters into the ascending and descending areas. Both of the ascending 'a' and descending 's' versions performed better than x-height versions. The PykeTest version of the descending 'f' also performed better than the non-descending forms. In the case of the SpencerTest and the PykeTest distance studies, and the the PykeTest exposure study, s3 showed a reliably better performance than the x-height s1, and a4 showed in general no statistically reliable differences compared to other two-storey 'a' versions. Implementing the high performing unfamiliar versions in a font within a new typeface would theoretically place the font on an equal legibility level to a font of familiar letterforms within the same typeface. The two fonts will have the same level of legibility but very different familiarity levels. Studying readers' experience with these different versions would be an interesting subject for future research into typeface familiarity.

The hypothesis that a single storey 'a' is less legible than the double storey 'a' was confirmed. With a high level of misreadings for 'o' and 'q', we recommend against the single storey 'a'.

We cannot conclude that creating differences between the letters 'n' and 'u' increases the legibility of a typeface. Neither a tailless form of the letter 'u' nor lowering the connection point of the letter 'n' had the intended effect. The recognition rates were comparable to the more common letterform.

The hypothesis that more open versions of 'c' and 'e' are more legible than more closed forms was not confirmed. While there was some indication that this was true for the letter 'e', there was no indication this was also true for the letter 'c'. The extremely closed e4 versions performed worse than the more open e1, e2, and e3. Except for the PykeTest at distance, the more closed c3 did not perform reliably worse than c1 or c2.

The study confirmed the notion that the performance of letter shapes varies according to the situa-

tion in which it is presented, and that some features are most important in distance viewing and others are most important in the parafoveal view. There are many differences between a letter from one typeface and the same letter in another typeface. The present method of studying within-font matters provides data that has a practical use for the design of new typefaces. We found that within a single typeface design that wide letters 'j', 'l', and 't' performed better than the narrow form in the same design. If we had instead compared a typeface like Courier with wide letter designs to a typeface like Helvetica with narrow letter designs, it would be much harder to reach the same conclusions because of the many inherent differences between Courier and Helvetica. Similarly, it is difficult to compare Futura's single storey 'a' to Times New Roman's double storey 'a'. This test compared the single and double storey 'a' within the same typeface and found the single storey 'a' to be less legible. This technique can help us reach a much stronger conclusion about the legibility of different typeface designs, and would be a useful technique to apply to the design of any new typeface.

## *Note*

**1.** This paper is partially based on the PhD thesis of the first author while affiliated with the Royal College of Art, London, UK.

## *References*

Bouma, H. (1971). Visual recognition of isolated lower-case letters. *Vision Research*, 11, 459–74.

Chaparro, B. S., Shaikh, A. D., & Chaparro (2006). Examining the legibility of two new clearType fonts [online]. *Usability News*, 8(1), available from http://www.surl.org [Accessed 22 October 2008].

Fox, D., Chaparro, B.S., & Merkle, E. (2007). Examining legibility of the letter "e" and number "o" using classification tree analysis [online]. *Usability News*, 9(2), available from http://www.surl.org [Accessed 22 October 2008].

Garvey, P.M., Pietrucha, M.T., & Meeker, D. (1997). Effects of font and capitalization on legibility of guide signs. *Transportation Research Record*, 1605, 73–79.

Geyer, L. H. (1977). Recognition and confusion of the lowercase alphabet. *Perception & Psychophysics*, 22(5), 487–90.

Harris, J. (1973). Confusions in letter recognition. *Professional Printer*, 17(2), 29–34.

Larson, K. (2005). The science of word recognition. *Typo*, vol. 13, (pp. 2–11).

Legros, L.A., & Grant, J.C. (1916). *Typographical printing surfaces: the technology and mechanism of their production.* London: Longmans, Green, and Co.

Lonsdale, M.d.S. (2007). Does typographic design of examination materials affect performance? *Information Design Journal*, 15(2), 114–139.

McClelland, J.L., & Johnston, J.C. (1977). 'The role of familiar units in perception of words and nonwords'. *Perception & Psychophysics*, 22(3), 249–261.

McClelland, J.L., & Rumelhart, D.E. (1981). An interactive activation model of context effects in letter perception: part 1. an account of basic findings. *Psychological Review*, 88(5), 375–407.

Ovink, G. W. (1938). *Legibility, atmosphere-value, and forms of printing types*. Leiden.

Paap, K.R., & Noel, R.W. (1991). Dual-route models of print and sound: still a good horse race. *Psychological Research*, 53, 13–24.

Pelli, D.G., & Tillman, K.A. (2007). 'Parts, wholes, and context in reading: a triple dissociation' [online]. PLoS ONE 2(8): e680, New York University. Available from: http://psych.nyu.edu/pelli [Accessed 16. November 2008].

Pyke, R. L. (1926). *Report on the Legibility of Print*. Medical Research Council, London: His Majesty's Stationery Office.

Rayner, K. (1978). Foveal and parafoveal cues in reading. In J. Requin (ed), *Attention and performance VII*. Hillsdale, NJ: Erlbaum.

Rayner, K. & Pollatsek, A. (1989). *The Psychology of Reading.* Englewood Cliffs, NJ:Prentice-Hall International.

Rayner, K., McConkie, G.W., & Ehrlich, S. (1978). Eye movements and integrating information across fixations. *Journal of Experimental Psychology: Human Perception and Performance*, 4, 529–44.

Rumelhart, D.E., & McClelland, J. L. (1982). An interactive activation model of context effects in letter perception: part 2. The contextual enhancement effect and some tests and extensions of the model. *Psychology Review*, 89(1), 60–94.

Sasson, R. (2001). 'Through the eyes of a child: perception and type design'. in: D. Jury, (ed), *Typographic Writing*, ISTD.
Sheedy, J.E., Subbaram, M.V., Zimmerman, A.B., Hayes, J.R. (2005). Text legibility and the letter superiority effect. *Human Factors*, 47, 797–815.
Tinker, M.A. (1964). *Legibility of print*. U.S.A.: Iowa State University Press.
Tracy, W. (1986). *Letters of credit: a view of type design*. Boston: David R. Godine.
Waller R. (2007). Comparing typefaces for airport signs. *Information Design Journal*, 15(1), 1–15.

## *About the Authors*

**Sofie Beier** is a designer, researcher and lecturer employed at The Danish Design School. She has a PhD from the Royal College of Art in London, on the subject of typeface familiarity and its relation to legibility. In the study of legibility related matters, her academic research focuses on integrating design approaches with methods applied by the scientific communities.
Her typefaces are published by T26, Die Gestalten Verlag, and FontShop.

Email: sbe@dkds.dk

**Kevin Larson** received his PhD in cognitive psychology in 2000 from the University of Texas at Austin. His academic research was on word recognition and reading acquisition. His passion is understanding the impact of typography on the reading experience, and applying that understanding towards improving the on-screen reading experience. He is a member of Microsoft's Advanced Reading Technologies team.

Email: kevlar@microsoft.com

## *Contact*

The Danish Design School
Strandboulevarden 47
2100 Copenhagen Ø
Denmark

Microsoft Advanced Reading Technologies
1 Microsoft Way,
Redmond, WA 98052
USA