Chapter 3

Sitka: a collaboration between type design and science

Kevin Larson and Matthew Carter

We recognize words by first recognizing individual letters, then using the letters to build a word [Larson, 2004; Rayner *et.al*, 2012]. Words become more readable by making each of the individual letters more recognizable. This chapter is about the development process for a new typeface named Sitka. During the typeface's development, we tested how well people could read each of the letters in the typeface, and used the test results to inform design decisions. While the test results needed to be applied conscientiously, we discovered that typeface design could be successfully integrated with scientific legibility testing.

1.1 The design brief

A number of requirements for this new typeface were established at the outset. The design was to be a general-purpose typeface primarily for use on screen, a serif design with a family of Roman, Italic and Bold (Bold Italic was added later). Wide Latin language support was essential, with Greek and Cyrillic included. Figure 1 shows a sample of the lowercase Latin letters.

The quick brown fox jumps over the lazy dog

Figure 1: The famous pangram set in the typeface Sitka.

The typeface had to be optically scaled, in other words it would have different versions, each one optimized for setting at a specified output size or narrow range of sizes. Most of the attention would go into the extremes, particularly the smallest size. The number of intermediate masters required to cover the entire range of sizes, from 6 point to 36 point and above, was not decided until late in the development.

One aspect that was fundamental to the design process, and highly unusual in the practice of type design, was the decision to test the legibility of the design as it progressed. The design was tested at each stage, the results analyzed, and used to improve the next stage. This iterative process was a collaboration between designers and scientists. The aim was to design a typeface as legible as we could make it, and in the course of doing that, to learn more about legibility and the testing of legibility in general. We hoped that what we learned would be helpful beyond this particular exercise.

1.2 The design process

Thirteen tests during the development of Sitka examined many aspects of the typeface: there were tests focused on the lowercase Roman, uppercase Roman, lowercase italic, lowercase Greek, lowercase Cyrillic, and the lowercase large optical size. The studies used a time threshold letter recognition measure to compare the current version of the font to either an earlier version of the font or to a close comparison font.

A massive amount of data was collected over the course of this design process, too much to discuss here. Instead we will discuss some of the more interesting findings and how we came to our design decisions. The findings include: it's not possible to test everything; a large x-height comes with a cost; a large x-height harms Greek letters; letter width and letter frequency impact letter recognition; it's harder to recognize a letter when next to other letters; large size designs are optimized for elegance not legibility; a typeface is a beautiful collection of letters not a collection of beautiful letters; and sometimes it's necessary to ignore the test results entirely.

1.3 It's not possible to test everything

A goal of this project was to run the studies quickly so design iterations could happen quickly. Each study looked for improvements between subtly different letterforms with a minimal number of study participants (usually 10-12) in order to get results as quickly as possible. We could have tested more people to be more confident of the results, but we decided that it would be better to run more tests with fewer people. As a trade-off we decided it was ok to make decisions with imperfect data. In order to get the most consistent data from each reader we chose to include only people who were 18-38 years old, native readers of English (native Greek and Russian readers for the Greek and Cyrillic studies), and had either normal or corrected-to-normal vision.

Readers were placed 150cm from the high DPI screen (144 DPI), and the letters were 36 points tall. This is three times further than normal reading distance and three times larger than typical text size. This kept the visual angle typical for reading, while allowing us to use large sized letters to test the design without any of the artifacts that can occur when there are too few pixels to represent the design properly.

The reader starts a single trial by looking at a mark on screen, followed by the test material that appeared for only part of a second. The test material was either a single letter displayed at the location of the mark, or was a sequence of three letters with the middle letter centered at the location of the mark. We called the three-letter sequence the context condition as each letter was placed between the letters that it most commonly appears between in written English, though we ensured that the three letters was never a word. In either case, the reader only needed to recognize the letter that appeared at the mark. After the test letters appeared briefly, they were replaced by a letter mask that limited any further processing of the test. When the trial was over, the reader was asked to identify the letter at the mark by pressing that letter on the keyboard. After each trial the reader was told if they identified the letter correctly.

At the beginning of each session a staircase procedure was used to find the fastest presentation time at which a reader could achieve a 50% accuracy rate. It was important for readers to make some errors to detect accuracy differences between letters. If the presentation times are long, even a poor typeface could have an accuracy rate of 100%. The presentation time was decreased by 17ms with each correct response and increased by 17ms with each incorrect response until a stable 50% accuracy was reached over 156 trials.

The testing method was refined over the first couple of studies. As an example, in the first study we compared two designs of the 26 Roman lowercase letters that differed in the openness of some of the letterforms. Under our initial context condition and kind of mask, we found an anomalous number of misrecognitions of the letter n. This didn't make sense; we feared that either flanking every test letter with n, o, or x or the mask was causing this effect. After changing our flanking letters to frequently occurring letter combinations and changing our mask to random letters, we no longer found surprising numbers of misrecognitions for any particular letter.

Another refinement we made during the earliest studies was to change the number of letters we investigated per study. The results from the first study of only 52 letters (26 lowercase letters in two different fonts) were very stable. Letters that were clearly different in design between the two fonts showed differences of only a few percent when tested. For the second study we tried testing many more conditions at once. We investigated the lowercase, uppercase and numbers in three different fonts. The increase from 52 letters to 186 letters led to very unstable results. In some cases there were improbable 20% accuracy differences between letters that were nearly identical.

The method was stable after fixing some of these early problems through trial and error. A tradeoff that we made was to not test every letter; we carefully choose only the most important letters to test. This left many omissions in the testing including extended character sets, uppercase italic letters, and the entire bold weight.

1.4 A large x-height comes with a cost

A commonly held truth among type designers is that legibility is improved by a large x-height [Beier and Dyson, 2014]. The term x-height is commonly used to mean the ratio between the neutral-height letters (e.g. lowercase x) and the overall vertical dimension of the typeface measured from the top of ascending strokes to the bottom of descending strokes. Verdana (Figure 2) is an example of a typeface that has a large x-height and is recognized as being very legible because of it. Sheedy, Tai, Hayes, and Preston [2006] found that Verdana was more legible than an entire suite of highly-touted brand new type designs.

Figure 2: A comparison of x-height proportions of Futura, Times New Roman, Verdana, and Sitka. All have been scaled to the same ascender height. Futura on the left has small x-height while Verdana and Sitka on the right have large x-heights.

We found that a large x-height comes with a trade-off. While the large size helps neutral height letters, it hurts the ascending and descending letters. Table 1 shows the accuracies for the lowercase letters in the fourth round (labeled before) of testing. The neutral height letters (acemrorsuvwxz) are recognized 48% correctly while the ascending (bdfhiklt) and descending (gjpqy) letters are recognized 38% and 33% correctly respectively.

Figure 3: The descending features are longer in the after (right) version of Sitka.

After discovering the trade-off, we tried slightly increasing the length of the descenders without changing the height of the neutral or ascending letters (Figure 3). This had the effect of slightly increasing the accuracy of the neutral-height letters from 48% to 50%, and the ascending letters from 38% to 39%. The descending letters showed the largest increase from 33% to 37% accuracy (Table 1).

In order to accommodate the longer descenders we planned at first to increase the overall height of the face with the effect of reducing the amount of vertical clearance between successive lines of text. At a later stage we decided instead to retain the overall size of the face which meant a proportional reduction in the height of both neutral and ascending letters. Because lengthening the descenders had proved to benefit not only the descenders but, unexpectedly, the neutral and ascending letters as well we considered extending the ascenders and descenders still further, but we feared this would have come at a cost to the x-height. In the end we abandoned this idea because the neutral height letters contain the most frequently occurring letters in English.

Table 1: Letter recognition accuracies for two versions of the Sitka typeface. The largest change from before to after was a lengthening of the descenders.

	Before	After		Before	After		Before	After
a	56%	54%	j	26%	28%	s	56%	59%
b	35%	33%	k	43%	40%	t	38%	39%
c	30%	34%	l	39%	41%	u	54%	51%
d	26%	33%	m	63%	64%	v	35%	36%
e	71%	75%	n	36%	31%	w	53%	54%
f	43%	46%	o	62%	59%	x	36%	46%
g	35%	41%	p	41%	43%	y	33%	41%
h	40%	41%	q	31%	32%	z	34%	38%
i	40%	38%	r	44%	45%	total	42%	44%

1.5 A large x-height harms Greek letters

The x-height decision for Latin letters determined the x-height for the Greek and Cyrillic scripts. While it is possible to choose different x-heights for different writing systems, having consistent x-heights improves the harmony when scripts are combined in the same document, such as Greek with German, or Russian with English.

When testing the Greek lowercase, we closely followed the procedure that we established for the English tests. While the tests took place in the United States, all of the readers were both native speakers and readers of Greek. We compared letter recognition of the Greek letters in Sitka to the Greek letters in Georgia (Figure 4), the most similar typeface for a comparison in a first study. As with English, the three-letter sequences were chosen for being the most common in Greek, without being a word.

While the neutral height letters in both typefaces were recognized correctly 45% of the time, Table 2 shows that Georgia performed better for both the ascending and descending letters. The ascending letters (δθλ) were recognized 52% correctly in Sitka to 62% in Georgia, and the de-

Figure 4: Georgia Greek (left) and Sitka Greek (right). The x-heights have been scaled to be the same, to show the relatively taller proportions of the ascending letter in Georgia.

scending letters (γημρχ) were recognized 40% correctly in Sitka to 44% in Georgia. Despite the advantages of the Georgia x-height for the Greek letters (Figure 4), we choose to retain the Latin x-height in order to better harmonize across different scripts. If Sitka was designed for Greek only, or for Greek first (with harmonizing Latin and Cyrillic scripts to follow), we would have designed a smaller x-height ratio.

Table 2: Letter recognition accuracies for each Greek letter in the typefaces Georgia and Sitka. The descending letters (γημρχ) perform relatively poorer in Sitka than in Georgia.

	Georgia	**Sitka**		**Georgia**	**Sitka**		**Georgia**	**Sitka**
α	51%	48%	ι	41%	41%	ρ	43%	41%
ά	40%	35%	ί	38%	31%	σ	42%	40%
β	67%	61%	κ	52%	49%	τ	45%	41%
γ	41%	37%	λ	57%	50%	υ	32%	32%
δ	66%	60%	μ	42%	37%	ύ	29%	31%
ε	53%	49%	ν	37%	33%	φ	53%	57%
έ	52%	36%	ξ	53%	47%	χ	56%	46%
ζ	58%	49%	ο	46%	44%	ψ	46%	41%
η	40%	37%	ό	41%	34%	ω	49%	55%
ή	29%	28%	π	53%	59%	ώ	40%	35%
θ	62%	47%						

1.6 The entire alphabet should be made of the letter m

There are aspects of letter design which have a big impact, but we cannot control. The first should be obvious, though we haven't found any explicit references to it in the literature. Letters that occur more frequently are more likely to be recognized correctly than letters that are less frequent. This kind of frequency effect in words is perhaps the most robust finding in all of reading psychology. Frequently occurring words are recognized faster than less frequently occurring words. The same appears to be true for letters. There is a very strong correlation (r =.58) between the frequency of letters in English and the accuracy that we recorded in our studies (Table 3). The letter best recognized in Sitka was the letter e, and it is by far the most frequent letter in English.

Table 3: English letter frequency correlates with Sitka letter recognition accuracy

	Frequency	Accuracy		Frequency	Accuracy		Frequency	Accuracy
e	12.7%	74%	d	4.3%	42%	p	1.9%	49%
t	9.1%	46%	l	4.0%	48%	b	1.5%	49%
a	8.2%	59%	c	2.8%	42%	v	1.0%	41%
o	7.5%	66%	u	2.8%	61%	k	0.8%	44%
i	7.0%	46%	m	2.4%	70%	j	0.2%	30%
n	6.7%	44%	w	2.4%	59%	x	0.2%	43%
s	6.3%	62%	f	2.2%	47%	q	0.1%	36%
h	6.1%	52%	g	2.0%	40%	z	0.1%	44%
r	6.0%	50%	y	2.0%	37%			

Another aspect of letter design with a big impact is the width of a letter. Wide letters tend to perform better than narrow letters. Beier and Larson [2010] investigated the well-known issue that certain narrow letters (ijlt) are very difficult to distinguish. They found that designing these letters to be wider helped to improve their recognition performance. This research influenced the Sitka design: the narrow letters were made as wide as feasible. Even with this design influence, there is a very strong correlation (r =.42) between the width of a letter and its recognition accuracy (Table 4). We're certainly not the first to recognize that narrow letters are difficult to recognize. The letter m is the second best recognized letter in Sitka, and it is the widest lowercase letter.

To examine the problem of narrow letters we looked not only at accuracy rates, but also at specific misrecognitions. Figure 5 is a complex visualization, but full of interesting data. The first four columns show the mis-recognitions for the letter f. The first column is the letter f when shown in isolation, the second, third, and fourth columns for the letter f when flanked by aft, efo, and ife. There were no misrecognitions for the letter f when it was presented in isolation or when it was flanked with efo. When f was flanked with aft it was frequently misrecognized as the letter k, (visualized as a ¾ full circle). When f was flanked by ife it was very frequently (full circle) misrecognized as the letter k, and sometimes (1/4 circle) as the letter h. The next 16 columns show the same kind of visualizations for the letters i, j, l, and t.

Table 4: The width of each Sitka letter in design units correlates with letter recognition accuracy

	Width	**Accuracy**		**Width**	**Accuracy**		**Width**	**Accuracy**
m	1737	70%	k	1034	44%	c	854	42%
w	1542	59%	o	1030	66%	s	831	62%
n	1149	44%	g	1020	40%	r	830	50%
h	1128	52%	x	998	43%	t	659	46%
u	1120	61%	v	993	41%	f	623	47%
p	1096	49%	y	971	37%	i	537	46%
d	1085	42%	a	964	59%	j	524	30%
b	1065	49%	e	924	74%	l	520	48%
q	1065	36%	z	883	44%			

The letter f is most likely to be misrecognized for the letter k, but only when flanked by other letters. The letter i is likely to be misrecognized for l and j, but only when in isolation. The letter j is more likely to be misrecognized when flanked by other letters, and is misrecognized for f and t as well as for the more surprising g and h. The letter l is more likely to be misrecognized for other ascending letters h and k when flanked by other letters. The letter t is strongly misrecognized for the letter k when flanked by other letters. The fact that narrow letters are often misrecognized as wider letters when flanked is a very interesting finding. It shows that neighboring letters interact with each other and that those interactions need to be considered in more detail.

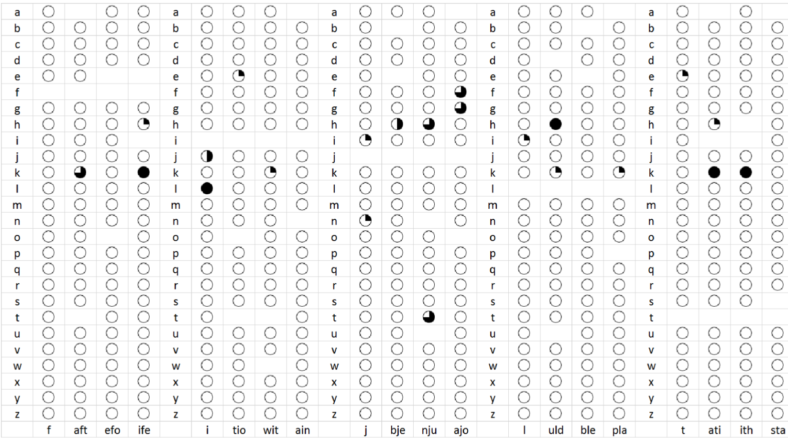

Figure 5: Misrecognitions for the narrow letters in study 4

1.7 It's harder to recognize a letter when next to other letters

Most words have more than one letter. This greatly complicates the measurement of letter recognition accuracy. It was discovered more than 100 years ago that letters are recognized more accurately when quickly presented as part of a word than when presented in isolation. The Word Superiority Effect comes from our top-down knowledge of language and tells us little about the bottom-up qualities of type design [Larson, 2004]. This effect is contrasted with the more recently discovered Letter Superiority Effect. When letter recognition is size constrained rather than time constrained, letters contained within words needed to be 10-20% larger in order to reach the same level of accuracy as letters in isolation [Sheedy *et. al.*, 2005].

We avoided the Word Superiority Effect by testing each letter in the context of common letter sequences that weren't complete words, in addition to testing each letter in isolation. Without words, our data was more similar to the Letter Superiority Effect. On average letters presented in isolation were recognized 58% correctly compared to 45% correctly for letters in the context of letter sequences.

While letters in isolation were recognized more accurately, most often a letter design that performed better in isolation also performed better when there were flanking letters. In one study where we compared a more open design to a more closed design there were differences between which aacceess3399

Figure 6: The more open letters (left) performed better when flanked by other letters, while the more closed letters (right) performed better in isolation.

design performed best in isolation and when there were flanking letters. Figure 6 shows the more open design on the left, where the aperture between the inside and outside of each letter is larger, and the more closed design on the right, where the aperture between the inside and outside of each letter is smaller. The more closed design performed better with the letters in isolation while the more open design performed better when the letters were flanked. Because it is far more common to recognize letters within a word, we choose to move towards the more open design than the more closed design.

When a misrecognition occurred for a letter with flanking letters, more than 26% of the misrecognition were for the flanking letter to the left while only 10% were for the flanking letter to the right. We speculate that the left flanking letter was more likely to be reported because of the left-to-right reading direction in the orthographies that we tested, but we would need to confirm that the opposite happens in a right-to-left language to be confident in that conclusion.

Because of the frequent misrecognitions with flanking letters, we considered what we could do to reduce these confusions. Perea and Gomez [2012] showed that word fixation times in continuous reading are faster when there is more space between letters. In more than one iteration of the design we increased the inter-letter spacing to reduce flanking letter confusions with only modest success. We went only so far with this idea because if the space between letters becomes too big the word falls apart.

1.8 The large size-specific designs are optimized for elegance (The Berlow-Hudson Hypothesis)

Historically, typefaces were designed differently for large sizes than for small sizes. This is sometimes called optical scaling or size-specific design. A typeface designed for small size will have a taller x-height, wider letters, less difference between thick and thin strokes, and wider spacing than a typeface designed for larger sizes. In metal type each size had to be made separately. In digital type the same letter can be displayed at any size. The advantage to digital type production is the ease and speed of designing a single version of each letter, but what is lost is the size-specific tuning. Sitka is one of several dozen digital typefaces that have different outlines optimized for different output sizes [Ahrens and Mugikura, 2013].

For the legibility testing aspects of this project we focused our attention on the Sitka Small size because we saw its optimization as having the greatest effect on legibility. In agreement with Harry Carter [1937/1984], we felt that the larger sizes were optimized for elegance and visual interest: 'Shortened descending and ascending strokes are unforgivable on bodies over 18-point. It is quite legitimate to shorten the tails of the small founts to increase legibility and to lengthen them in the display sizes of the same face for the sake of elegance.' An alternative hypothesis that was disputably proposed by David Berlow (personal conversation), and later taken up by John Hudson (personal conversation), claimed that the size-specific adjustments for larger sizes are in fact legibility optimizations for larger text.

To test which hypothesis is correct we used the same letter recognition method, but because typical reading situations were important to the hypothesis we shortened the testing distance in the study to a very typical 50cm distance. We tested a small and large size-specific adjustment, Sitka Small and Sitka Banner respectively (Figure 7), each at two physical sizes, 9 point and 36 point. We did not control for x-height as the x-height difference is one of the key variables that we tested, an element that is fundamental to size-specific adjustments. During piloting of the test we discovered that we could not use the same presentation time for both sizes. 36 point letters are easier to recognize than 9 point letters, so two different calibration times were used. If larger size-specific designs are optimized for legibility than we would expect Sitka Small to perform better at 9 point and expect Sitka Banner to perform better at 36 point. If larger size-specific designs are optimized for elegance than we would expect Sitka Small to perform better both at 9 point and 36 point.

Sitka Small performed better than Sitka Banner at both 9 point and at 36 point. Accuracy at 9 point was 48% for Sitka Small to 39% for Sitka Banner. Accuracy at 36 point was 56% for Sitka Small to 49% for Sitka Banner. The advantage for Sitka Small was statistically reliable, $F=29.3$, p=.0001, and seen for every study participant, and for 25 of 26 letters. The difference between sizes is unfair to compare because we recalibrated the presentation time to present the 36 point size faster. Even with the faster speeds, both Sitka Small and Sitka Banner had higher accuracy rates at 36 point. If we ignore the recalibration and run the statistical test anyway, we find that the size difference is reliable, F=7.0, p=.02, and that there is no interaction between size and font, F=.2, p=.66.

This strongly indicates that the size-specific adjustments made for large sizes do not increase legibility for large sized text. If we want increased legibility at large sizes, we are better served using a small size-specific design. If our goal is instead some level of elegance or personality, then a large size-specific design is appropriate.

The quick brown fox jumps over the lazy dog

The quick brown fox jumps over the lazy dog

The quick brown fox jumps over the lazy dog

The quick brown fox jumps over the lazy dog

The quick brown fox jumps over the lazy dog

The quick brown fox jumps over the lazy dog

Figure 7: Size-specific adjustments in Sitka: Sitka Banner, Sitka Display, Sitka Heading, Sitka Subheading, Sitka Text, and Sitka Small

1.9 A typeface is a beautiful collection of letters, not a collection of beautiful letters

At the start of the project we analyzed a wide variety of existing research into making legible letters. We were strongly influenced by Sheedy, Tai, Hayes, and Preston's [2006] research on the legibility of individual letters in a dozen top typefaces. Each letter's distance threshold was measured which gave us information about the relative legibility of very different styles of typefaces. One finding of this research was that the best performing typeface for one letter wasn't necessarily the best performing typeface for another letter. Each letter is a unique design. For example the typeface Centaur had the top performing letter s, while Verdana had the top performing letter a, and DIN had the top performing letter m (Figure 8).

An extreme argument could be made that we could create a new typeface by combining the top performing letters into a new Frankenfont. Throughout the project we were cognizant that we wanted to create a typeface and not a collection of letters. We used the earlier studies and our own tests as inspiration for characteristics for each letter that could improve their legibility within a coherently designed typeface.

One example of considered design came from a study of terminals. We measured the performance of the typeface with both teardrop and flared terminals and found that the letters c, f, and j performed 6% better on average with the flared terminal, while the letter a performed slightly better with the teardrop terminal (Figure 9). We could have made the decision to use the flared terminal with c, f, and j, while using the teardrop terminal with a. This inconsistency would have favored the parts over the whole, and resulted in a collection of letters rather than a typeface. We decided to use the flared terminal throughout the design.

Figure 8: Frankenfont of the Centaur s, Verdana a, and DIN m.

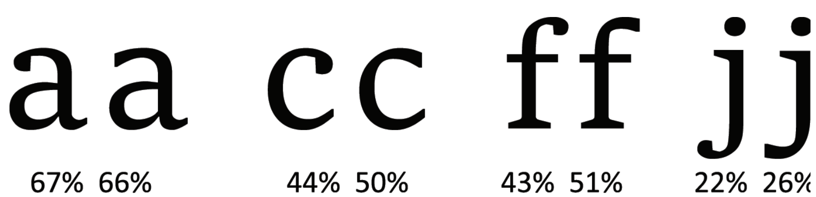

Figure 9: Letter recognition accuracy for four letters with teardrop terminals (left) and with flared terminals (right).

1.10 There can be considerations other than test results

The goal of the legibility research was to provide useful data to influence the design, not to mindlessly determine the design. One example came from the Greek study mentioned earlier in which native Greek readers compared letter recognition between Georgia and Sitka. We tested all of the base letters as well as the seven vowels that can have tonos marks in modern monotonic Greek. Prior to the study we had heard feedback from

Greek type design experts and from Greek readers that the tonos marks in Georgia are a particularly terrible design. For Sitka we designed a tonos mark that experts said was a better and more traditional angled form. We expected the angled design to perform better than the upright tonos mark.

The letter recognition for the unmarked vowels were pretty similar between Georgia and Sitka. 45% of the trials were recognized correctly with Georgia and 44% were recognized correctly with Sitka. The interesting comparison though was how the accuracy changed with the addition of the tonos mark. When the upright tonos mark is added to the Georgia vowels, the recognition accuracy decreased slightly from 45% to 44%. But when the angled tonos mark was added to the Sitka vowels, the recognition accuracy decreased dramatically from 44% correct down to 33% correct (Figure 10).

The results are compelling that Greek readers are more successful at recognizing the upright Georgia tonos marks than the angled tonos mark in Sitka. This left us with the decision between changing the Sitka tonos mark to a more-legible, upright design or staying with the less-legible, angled tonos mark in deference to the advice of experts in contemporary Greek orthography. In this case we decided that it was more important to go with the opinions of experts and readers.

1.11 Conclusions

The novel collaborative process that was used in the development of the Sitka typeface family worked well in the sense that the participants, designers on the one side, scientists on the other, saw very much eye to eye. It never happened that designers produced something that the scientists said was untestable, or the scientists produced results that the designers said were unusable. The process was always a matter of discussion and interpretation between participants who had great confidence in one another's abilities.

Some of the test results were anticipated, such as the problems inherent in narrow letters. In such a case our findings tended to reinforce the conclusions, or suggestions, of earlier researchers. The corollary, that lowercase m is an excellent letter, had perhaps not been so well established, and the superior legibility of the most frequent letters in the language seemed to be a discovery, if not a very surprising one. Despite our best efforts cer-

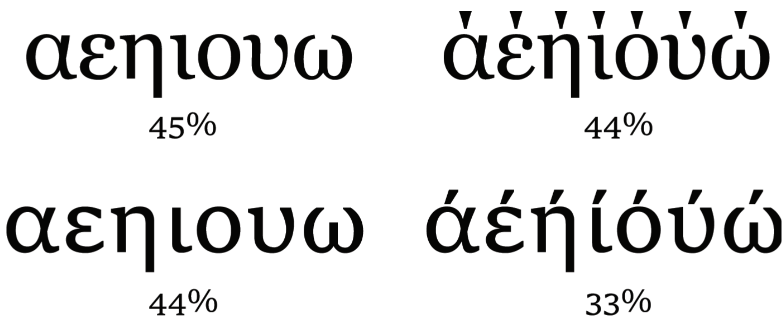

Figure 10: Letter recognition accuracy for vowels with (right) and without (left) tonos marks for both Georgia (top) and Sitka (bottom).

tain problems with the Latin alphabet remain intractable; there are practical limits to how wide an f can be stretched, or how long a descender can extend, and it's a hopeless task to try to persuade languages to dispense altogether with diacritics.

Although the type designer will never solve these problems, it is very important to understand the reasons behind them and to mitigate them to the extent that it's possible.

All of us who took part in this project felt that Sitka as it was completed and released was a significant improvement in legibility. We also felt that the design improved dramatically over the course of the development effort. Tantalizingly, we did not demonstrate statistically reliable reading speed benefits as a result of the improvements. We learned a great deal from this project, but it must be said that there is much about the legibility of type and its measurement that is not yet fully understood.

1.12 Acknowledgements

The design and development of Sitka was a team effort. It could not have been completed without Geraldine Wade and Michael Duggan, the invaluable project leaders at the beginning and end of the project respectively. We would like to thank the other members of Microsoft's Advanced Reading Technologies, Greg Hitchcock, Paul Linnerud, Tanya Matskewich, and Rob McKaughan for analyzing the study results and so much more, Simon Daniels and Ali Basit our partners on the Microsoft Typography team, Kris Keeker and Libby Hanna for help in conducting the studies, John Hudson and Ross Mills for help on mastering the typeface, Gerry Leonidas for consulting on the Greek design and Greek studies, and Steven Sinofsky for funding the project. The Sitka typeface is dedicated to the memory of Bill Hill. Bill's passion and vision for Reading 2.0 live on in this typeface.

1.13 References

Ahrens, T. and Mugikura, S. (2013). *Size specific adjustments to type designs*. (Just Another Foundry).

Beier, S. (2012). *Reading letters: designing for legibility*. (BIS Publishers).

Beier, S. and Dyson, M.C. (2014). The influence of serifs on 'h' and 'i': Useful knowledge from design-led scientific research. *Visible Language*, 47(3), pp. 74-95.

Beier, S. and Larson, K. (2010). Design improvements for frequently misrecognized letters. *Information Design Journal*, 18(2), pp. 118-137.

Carter, H, (1984). Optical scale in type founding. *Printing Historical Society Bulletin*, 13, 144-148. Reprinted from Carter, H. (1937). Optical scale in type founding. *Typog-raphy*, 4.

Legge, G.E. and Bigelow, C.A. (2011). Does print size matter for reading? A review of findings from vision science and typography. *Journal of Vision*, 11(5), pp. 1-22.

Larson, K. (2004). The science of word recognition or how I learned to stop worrying and love the bouma. http://microsoft.com/typography/ctfonts/wordrecognition.aspx

Perea, M. and Gomez, P. (2012). Subtle increases in interletter spacing facilitate the encoding of words during normal reading. *PLoS ONE* 7(10): e47568.

Rayner, K., Pollatsek, A., Ashby, J. and Clifton, C. (2012). *Psychology of Reading*, Second Edition, (Psychology Press).

Roethlein, B.E. (1912). The relative legibility of different faces of printing types. *American Journal of Psychology*, 23(1), pp. 1-36.

Sheedy, J.E., Subbaram, M.V., Zimmerman, A.B., and Hayes, J.R. (2005). Text legibility and the letter superiority effect. *Human Factors*, 47(4), pp. 797-815.

Sheedy, J.E., Tai, Y-C., and Hayes, J.R. (2006). Individual character legibility. http://www.pacificu.edu/sites/default/files/documents/Individualcharacterlegibility.pdf