Acta Psychologica 228 (2022) 103623

Contents lists available at ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/locate/actpsy

The dispute about sans serif versus serif fonts: An interaction between the variables of serif and stroke contrast[☆]

Katsumi Minakata, Sofie Beier [*]

The Royal Danish Academy – Centre for Visibility Design, Denmark

ARTICLE INFO

Keywords:
Serif
Sans serif
Font
Stroke contrast
Word identification
Psychophysics

ABSTRACT

Aim: It is a long-lasting dispute whether serif or sans serif fonts are more legible. However, different fonts vary on numerous visual parameters, not just serifs. We investigated whether a difference in word identification can be attributed to the presence or absence of serifs or to the contrast of the letter stroke.

Method: Participants performed a word-recognition two-interval, forced-choice task (Exp. 1) and a classic lexical decision task (Exp. 2). In both experiments the word stimuli were set with four new fonts, which were developed to isolate the stylistic features of serif and letter-stroke contrast. Two measures (i.e., font-size threshold & sensitivity) were analysed.

Results: The threshold measure of both experiments yielded a single significant main effect of stroke contrast such that low stroke contrast elicited lower than high stroke contrast. The sensitivity measure of Experiment 1 yielded a single significant effect of the interaction between serifs and stroke contrast. Specifically, at the sans-serif level, low stroke contrast revealed better sensitivity, relative to high stroke contrast. At the serif level, the opposite stroke contrast pattern was observed.

Conclusion: Sans serif fonts with low stroke contrast yield better performance and if a serif font is used, high stroke contrast yields better performance than low stroke contrast. Limitations and future directions are discussed.

1. Introduction

Discussions about font choices often revolve around a focus on the two categories of sans serif and serif fonts. Serif fonts have small ornamentations at the stroke endings that sans-serif fonts do not have. The typographic literature proposes several arguments in favour of serifs, among them the belief that serifs emphasize stroke endings (Unger, 2007), because the horizontal shape of serifs emphasizes the reading direction by supporting the movement of the eye from left to right (Beier, 2012). It is also believed that serifs help the reader distinguish the letters while also linking them together to form words (McLean, 1980). The case against serifs is that they are simply extra features added to the letter, while sans-serif fonts represent the essential letter form (Frutiger et al., 1980).

In the history of legibility research, serifs are one of the most disputed typographic features (for a review of the early literature; see Lund, 1999). Numerous experiments have compared performance of multiple fonts, including both serif and sans serif examples; however, when the fonts are bundled into serif versus sans serif categories, the results are generally inconclusive, because fonts within the categories do not show similar reading performance (Bernard et al., 2002; Boyarski et al., 1998; Sheedy et al., 2005). One exception is an investigation into visual acuity with word stimuli set in 33 different fonts, which found a small but significant effect in favour of lowercase sans-serif fonts in a comparison of the collective performance of sans serif and serif fonts (Garvey et al., 2016). Others claim to be able to draw conclusions based on comparing reading performance for only two fonts belonging to different font families (Beymer et al., 2008; Dogusoy et al., 2016).

The main limitation of these approaches is that fonts from different font families vary in terms of multiple variables besides serifs, for example, letter proportions, letter skeleton, letter weight and stroke contrast. Thus, in a comparison of two fonts from different families it is difficult to isolate the effect of serifs from the effects of other font variables.

These flaws in experimental design have also been pointed out by others, who instead conducted experiments that isolated the effect of

[☆] This work was supported by the Danish Council for Independent Research [grant number DFF – 7013-00039].

[*] Corresponding author.
E-mail address: sbe@kglakademi.dk (S. Beier).

https://doi.org/10.1016/j.actpsy.2022.103623
Received 11 December 2020; Received in revised form 17 May 2022; Accepted 19 May 2022

serifs (Akhmadeeva et al., 2012; Beier & Dyson, 2014; Moret-Tatay & Perea, 2011; Morris et al., 2002; Perea, 2013). Examples include the multiple experiments comparing reading performance of the two fonts Lucida and Lucida Sans (see Fig. 1), which are designed to only vary with regards to the presence or absence of serifs. One of these experiments measured the effect by using rapid serial visual presentation (RSVP, letter/words presented one at a time at the same location) and found that at small sizes, the sans-serif font (i.e., Lucida Sans) could be read faster than the serif font (i.e., Lucida) although the effect disappeared at large font sizes (Morris et al., 2002). However, similar results were not found in a study of single-letter recognition (at reading acuity limit) with a different font family, which found no difference between sans serif and serif fonts, but when looking at the subgroup of letters that contained serifs on vertical extremes ('l', 'b', 'h', 'n', 'u'), the study showed a positive effect for serifs (Beier & Dyson, 2014).

Another experiment that used the Lucida font family employed a lexical-decision (LD) task and found a small but significant advantage for a sans-serif font (Moret-Tatay & Perea, 2011), while yet another experiment found no effect of these fonts on eye-movement measures (Perea, 2013). The latter finding was supported by experiments measuring reading speed for sans serif and serif versions within other font families, which yielded no significant differences between sans serif and serif font styles (Akhmadeeva et al., 2012; Arditi & Cho, 2005a).

Sans serif and serif fonts vary not only in terms of the presence or absence of serifs but also in terms of stroke contrast. Compared to sans-serif fonts, serif fonts tend to have a greater difference between the thickest and thinnest parts of a letter's stroke, a feature referred to as stroke contrast (see Fig. 2). The research literature shows almost no interest in this typographic characteristic. Except for a recent study into stroke contrast in bold serif fonts, which found that hairline strokes lower letter recognition (Beier & Oderkerk, 2021), other studies concerned with the effects of font style have mainly looked into letter complexity (Beier et al., 2018; Bernard & Chung, 2011; Pelli et al., 2006) and letter boldness (Beier & Oderkerk, 2019; Bernard et al., 2013; Burmistrov et al., 2016; Chung & Bernard, 2018; Macaya & Perea, 2014; Pelli et al., 2006; Sheedy et al., 2005). The main aim of the present paper was to isolate the two typographic features of serifs and stroke contrast and investigate whether a given difference in reading performance between serif and sans-serif fonts is attributable to serifs or to stroke contrast, and in addition being able to isolate these two features.

2. Experiment 1

We employed the well-established, method of constant stimuli (MOCS), which is a psychophysical technique used to measure and estimate perceptual thresholds by randomly and constantly presenting stimuli at different intensities (Kingdom & Prins, 2010) to obtain font-size thresholds. A perceptual threshold is defined as the minimum amount of physical energy (e.g., level of light intensity) required to detect a stimulus 50% of the time and we designed an experiment such that a font-size threshold could be estimated for each font condition.

To measure and estimate thresholds, a two-interval, forced-choice (2IFC) task is used, which consists of two time-intervals and a target stimulus is randomly placed in one of the two intervals with a distractor/noise stimulus in the other interval (Kingdom & Prins, 2010). The required response is to indicate in which of the two time-intervals the target stimulus was perceived (e.g., a correctly spelled word). Given the nature of the 2IFC task, one can, theoretically, obtain a correct response when one simply guesses randomly (e.g., if an observer always indicates interval one contained the target stimulus, then 50% of the time a correct response is obtained by chance). However, the 2IFC task also allows one to implement signal detection theory analysis, which makes it possible to calculate an observer's discrimination sensitivity. (i.e., both response bias and performance are considered) when discerning between a target signal and a target-plus-noise signal.

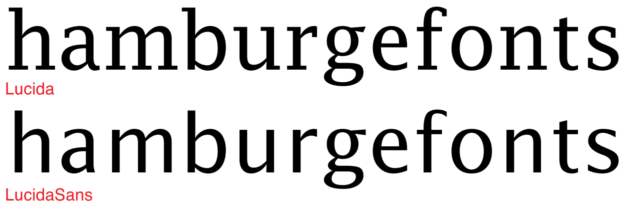


Fig. 1. Lucida varies solely in terms of the presence or absence of serifs. The font family of Lucida: Lucida (top) and Lucida Sans (bottom).

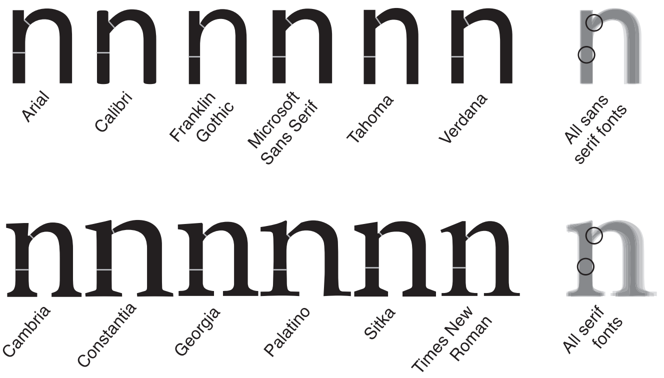


Fig. 2. A selection of the most used fonts in Windows 10. The top row shows sans-serif fonts; the bottom row shows serif fonts. By measuring the stroke contrast in the lowercase 'n' we found that the six sans-serif fonts have an average stroke contrast between thick/thin of 3/2, while the six serif fonts have an average stroke contrast of 3/0.8. Thus, the average stroke contrast in serif fonts is greater than the average stroke contrast in sans-serif fonts.

To measure and estimate font-size thresholds that are ecologically meaningful will give the opportunity to determine what font size is needed to identify a word. A classical, lexical-decision (LD) task involves the serial presentation of either a word or a pseudoword (i.e., pronounceable letter strings) and an observer is given the task to indicate if the stimulus was a word or a pseudoword (Meyer & Schvaneveldt, 1971). The reaction time and accuracy are recorded and analysed, which generally reveal that words elicit faster reaction times and higher accuracy rates relative to pseudowords – *word-superiority effect.*

It is suggested that these results measure one's ability to recognize words from pseudowords (i.e., not just stimulus detection). Therefore, Experiment 1 was designed to include a combination of a 2IFC and a word-recognition task, so that conclusions about the font-size thresholds could be drawn at the level of the lexical-identification stage within the human information processing stream. To our knowledge, the proposed experiment is the first to: (1) isolate typographical features of fonts; (2) utilize both an adaptive threshold estimation procedure and the MOCS; and (3) implement a 2IFC task with a word-recognition task.

Given that Moret-Tatay and Perea (2011) found evidence in a LD experiment that suggested sans-serif fonts yield better performance, relative to serif fonts, it was hypothesized that the sans serif conditions would elicit lower font-size thresholds, relative to the serif font conditions (H1). Additionally, the low stroke-contrast conditions were expected to yield lower font-size thresholds, relative to the high stroke-contrast conditions (H2). Finally, it is possible that these two factors will interact and change the pattern of results (H3).

2.1. Method

2.1.1. Participants

A total of 33 individuals were recruited via a website used for the recruitment of participants for studies (forsøgspersoner.dk). Thus, we collected a convenience sample with the following screening criteria: 18–40 year of age and normal or corrected-to-normal vision. There were

19 females (14 males), and the average age was 23 years (range: 18–39). All participants reported to have normal or corrected-to-normal vision and were paid (i.e., 48 USD) for 3 h of participation. All participants considered Danish their primary language. The research followed the tenets of the Declaration of Helsinki.

2.1.2. *Design*

A new font family was developed for this experiment (see Fig. 3). The experiment manipulated two repeated-measures, independent variables: (1) serif and (2) stroke contrast. The serif factor varied between two conditions: serifs being present or absent. The stroke contrast factor varied between two conditions: low or high stroke contrast. Stroke contrast was identical between serif and sans serif conditions and the brackets of the serifs were triangular with no curve. The high stroke contrast designs followed conventions of the Didot style fonts of vertical stress to the letters, while the low stroke contrast fonts followed conventions of slightly narrowing stroke-width in letter junctions (Beier, 2017).

2.1.3. *Apparatus*

An HP laptop with 8GB of RAM, Intel (R) Core i5-6300U 2.5 CPU GHz CPU and a 64-bit, Windows 10 operating system was utilized. The laptop monitor was an LCD with a 60 Hz refresh rate and had a resolution of 3000 by 2000 pixels. A second keyboard was used so that participants could be placed at a distance of two metres from the monitor. The scripts for the experiment were written in MATLAB in combination with the Psychophysics Toolbox and, thus, the stimuli were computer-controlled (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997). All the stimuli consisted of pairs of two-word categories; words and pseudoword (pronounceable orthographically legal word). Both word categories varied in length between three and six letters. The pseudoword had one or two central letters translocated. A total of 900 trials were included in the experiment and the same number of unique word-pairs were implemented (viz., all word- and pseudo-word pairs were novel). Nine font sizes were tested (each font size consisted of 20 trials) for each of the 4 experimental conditions. The stimuli were presented in white on a black background (i.e., 100% Michelson's contrast) in the centre of the screen with an exposure duration tailored to each participant based on an adaptive procedure (Watson & Pelli, 1983).

The QUEST algorithm is an adaptive Bayesian estimation procedure that can estimate a threshold value given a chosen psychometric function (e.g., cumulative normal distribution), slope, lapse rate, and guess rate. The lapse rate (i.e., incorrect response even though a perceptible stimulus was presented) and guess rate (determined by the task structure) were considered fixed parameters and were set to 0.019 and 0.50, respectively. Only two parameters (e.g., threshold and slope) were estimated and QUEST requires approximately 30–40 trials to obtain a reliable threshold estimate (Watson, 2017; Watson & Pelli, 1983). The inter-stimulus interval was 500 ms and the inter-trial interval was 1000 ms.

2.1.4. *Procedure*

Participants were greeted upon arrival, informed about the nature of the study, and asked to give their written consent in order to participate. After obtaining this informed consent, the researchers asked the participants to get as comfortable as possible using the chinrest, which was affixed to the table. The first step of the experiment was to obtain a threshold value for font size; this was done using the QUEST algorithm in combination with a word recognition task. Participants were exposed to one word-pair at a time, each pair consisted of a real word and a pseudoword. Throughout the QUEST procedure, the Helvetica font was used as a baseline font. Participants were informed that the target word stimulus would randomly appear in either the first or second interval and that they would complete the word recognition task under three different contexts, which were organized into two phases.

During phase one, the font-size, threshold estimation procedure was repeated three times and the exposure-duration, threshold estimation procedure was completed once. The font size varied on each trial based on a given participant's correct (or incorrect) responses and QUEST's statistical decision for the next stimulus intensity. The same procedure was implemented for the exposure-duration, threshold estimation.

During phase two, both the font size and font condition varied in a pseudo-random manner. The key-press responses to the word recognition task were recorded via a keyboard that featured a "left arrow" key to indicate a 'first interval' decision and a "right arrow" key indicated a "second interval" response. Each word was immediately backwards-masked by a salt-and-pepper noise patch with a 50 ms exposure duration to eliminate possible word afterimages due to neural or visual persistence (Sperling, 1965, see Fig. 4).

The QUEST-derived, font-size thresholds and their respective standard deviations were averaged to set the expected font-size threshold value for the experiment and to create the range of values for the font-size factor, which was necessary for the MOCS font-size threshold

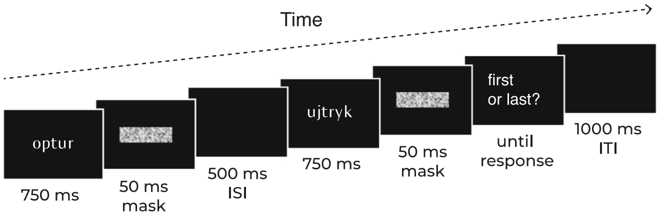
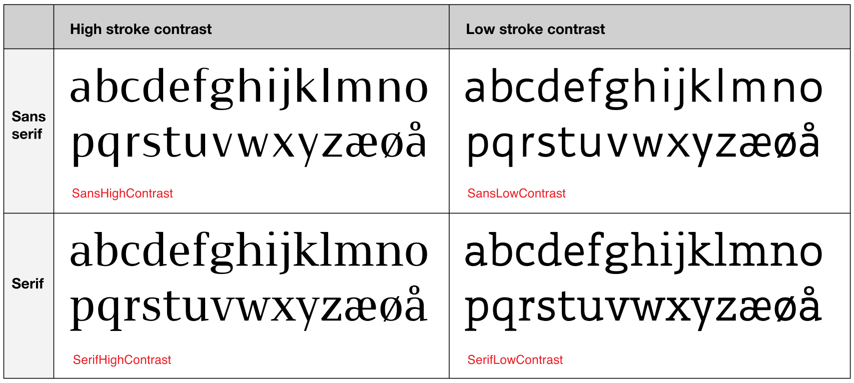


Fig. 4. Two interval forced choice task and word identification. For example, first a word is presented and then a pseudoword in serial presentation. Participants were asked to identify whether the word was shown first or last.



	High stroke contrast	Low stroke contrast
Sans serif	abcdefghijklmno pqrstuvwxyzæøå SansHighContrast	abcdefghijklmno pqrstuvwxyzæøå SansLowContrast
Serif	abcdefghijklmno pqrstuvwxyzæøå SerifHighContrast	abcdefghijklmno pqrstuvwxyzæøå SerifLowContrast

Fig. 3. The four font conditions developed for this experiment. The low stroke-contrast fonts have a ratio of thick/thin of 3/2.4, while the high stroke-contrast fonts have a ratio (thick/thin) of 3/0.8.

procedure. To create the font-size range, the average, font-size threshold served as the minuend and the average standard deviation was multiplied by two and served as the subtrahend, which created the lower limit of the font-size range. The same procedure was used to generate the upper limit of the font-size range except that the average standard deviation was multiplied by three. These two font-size values were used to create a linearly spaced font-size vector with nine elements that were the values of the font-size factor.

Experiment 1 (i.e., phase two) also incorporated a word-recognition task that was executed using the method of constant stimuli (MOCS), a psychophysical method wherein each stimulus pair is serially, constantly, and randomly presented to estimate an individual participant's perceptual threshold. The word recognition task was modified and designed to be a 2IFC task. There were a total of 720 trials whereby each font-size range of 9 elements contained 20 trials each (i.e., 180 trials per font condition); we tested four fonts. The experiment was split up into 10, equally sized, blocks (72 trials) and participants were able to take an optional resting break between each block. After the fifth block, the QUEST font-size threshold estimation task was repeated and, upon completion of the tenth block, was repeated once more (i.e., 5 QUEST thresholds).

2.1.5. Data analysis

RT was examined by comparing the RTs of correct responses (i.e., a word was in interval 1 and it was identified correctly) and incorrect responses (i.e., a word was in interval 2 and it was mistakenly identified as being in interval 1). Words elicited faster RTs relative to pseudowords, which means our 2IFC task was able to discern between words and nonwords regarding RT performance.

The dependent variables of interest were the participants' font-size threshold (α alpha) and sensitivity (d prime) for each experimental condition. These were obtained by fitting the percentage of correct responses, which were defined as a participant responding “first interval” when a word was located in the first interval. These nine font-size percentages were fitted with a cumulative-normal sigmoid function and a maximum likelihood procedure. To ensure the resultant font-size threshold values were more representative of the population parameter, the alpha parameter calculation assumed bias-free performance (20 trials per font-size condition), as opposed to a bias present assumption (10 trials per font-size condition). Because the participants completed a 2IFC task, the chance/guessing rate (γ gamma) was 0.50, the inattention/lapse-rate (λ lambda) was set to 0.019, and the threshold (α alpha) was considered the 75% point of the cumulative-normal function fit. These parameters were entered into the Psignifit Toolbox's maximum likelihood fitting procedure (Schütt et al., 2016). The alpha and beta values were allowed to be free parameter and the lambda and gamma values were fixed parameters (Prins & Kingdom, 2018).

Signal detection theory was implemented to assess the variations in participants' perceptual sensitivity. D-prime represents the participants' perceptual ability to distinguish between words and pseudowords (Macmillan & Creelman, 1991) and d prime values are normally distributed. The sensitivity measure was computed in the following manner: $d' = z(\mathrm{H}) - z(\mathrm{F})$ and $\beta = 0.5 \times [z(\mathrm{H}) + z(\mathrm{F})]$. H and F indicate the hit (correct detection of word in interval 1) and a false alarm (incorrect detection of word in interval 2) response rates and $z(p)$ indicates the inverse of the cumulative Normal distribution corresponding to response rate p. The calculation of d-prime assumed biased performance, which was corrected by taking both the H and F performance rates into account and was based on 10 trials per font-size condition. After the font-size thresholds were extracted, the font-size threshold value for each condition was used to obtain a d-prime value at participants' threshold level, which were, then, analysed with the following Bayesian framework.

Bayesian hierarchical linear models were fitted to both the font-size thresholds (*alpha*) and sensitivity values (d') as a function of the mean-centred factors, stroke contrast and serif, along with and their two-way interaction. The Stan modeling language (Carpenter et al., 2017) in R and the package *brms* (Bürkner, 2017; Stan Development Team, 2017; Stan Modeling Language, 2017) were utilized.

The models included maximal random-effect structures justified by the design (Barr et al., 2013), allowing the predictors of interest and their interactions to vary by participants. Both the mean for the stroke contrast reference cell (stroke contrast = low) and the mean for the serif type reference cell (serif type = sans), as well as their interaction (stroke contrast = low & serif type = sans) were given Gaussian priors (*alpha*: $\mu = 55$, $\sigma = 12$; d': $\mu = 1.8$, $\sigma = 2.5$). We used the *brms* package's default priors for standard deviations of random effects (a Student's *t*-distribution with $\nu = 3$, $\mu = 1.8$ and $\sigma = 2.5$), as well as for correlation coefficients in interaction models (LKJ $\eta = 1$).

Six sampling chains ran for 10,000 iterations with a warm-up period of 5000 iterations for each chain, thereby yielding 30,000 samples for each parameter tuple. For the marginal means and differences between them, we report the expected values under the posterior distribution and their 95% credible intervals (Cr. I.). For marginal mean differences, we also report the posterior probability that a difference δ is bigger than zero. If a hypothesis states that $\delta > 0$, then it would be considered *strong evidence* for this hypothesis would be if zero is not included in the 95% Cr. I. of δ and the posterior $P(\delta > 0)$ is close to one (by a reasonably clear margin). To extract the estimated marginal means from the posterior distribution of the fitted models we used the emmeans R package (Russell, 2021).

The probability of direction (pd) was used to determine whether the non-significant post hoc comparisons (i.e., sans serif low stroke contrast vs. serif high stroke contrast; sans serif high stroke contrast vs. serif low stroke contrast; & serif low stroke contrast vs. serif high stroke contrast) were equivalent (Makowski et al., 2019). The pd ranges from 50% to 100% and it represents the certainty regarding an effect's direction (e.g., positive or negative sign). The pd corresponds with frequentist *p*-values. A two-sided p-value of respectively 0.1, 0.05, 0.01 and 0.001 approximately corresponds to a *pd* of 95%, 97.5%, 99.5% and 99.95%. A low pd is related to no direction (no effect) and a high pd means there is a direction (positive effect). The “estimate_contrasts” function from the R package modelbased (Makowski et al., 2020) was applied to the brms model fit, which yielded the differences, 95% credible intervals, pd, and percentage in the ROPE.

The following model was estimated for both the font-size threshold and sensitivity dependent variables:

Level 1:

$$\mathrm{Font-size\ Threshold}_{\mathrm{ijk}} = \beta_{0\mathrm{j}} + \beta_{1\mathrm{j}}(\mathrm{Serif\ Type}) + \beta_{2\mathrm{j}}(\mathrm{Stroke\ Contrast}) + \beta_{3\mathrm{j}}(\mathrm{Serif\ Type})^{*}(\mathrm{Stroke\ Contrast}) + R_{\mathrm{ijk}}$$

Level 2:

$$\beta_{0\mathrm{j}} = \gamma_{00} + U_{0\mathrm{j}}$$

$$\beta_{1\mathrm{j}} = \gamma_{10}(\mathrm{Serif\ Type}) + U_{1\mathrm{jk}}$$

$$\beta_{2\mathrm{j}} = \gamma_{20}(\mathrm{Stroke\ Contrast}) + U_{2\mathrm{jk}}$$

$$\beta_{3\mathrm{j}} = \gamma_{30}(\mathrm{Serif\ Type})^{*}(\mathrm{Stroke\ Contrast}) + U_{3\mathrm{jk}}$$

Full Equation:

$$\begin{aligned}\mathrm{Font-size\ Threshold}_{\mathrm{ijk}} = {} & \gamma_{00}(\mathrm{Intercept}) + \gamma_{10}\left(\mathrm{Serif\ Type}_{\mathrm{ij}}\right) \\ & + \gamma_{20}\left(\mathrm{Stroke\ Contrast}_{\mathrm{ij}}\right) + \gamma_{30}\left(\mathrm{Serif\ Type}_{\mathrm{ij}}\right)^{*}\left(\mathrm{Stroke\ Contrast}_{\mathrm{ij}}\right) \\ & + U_{0\mathrm{j}}(\mathrm{Intercept}) + U_{1\mathrm{j}}\left(\mathrm{Serif\ Type}_{\mathrm{ij}}\right) + U_{2\mathrm{j}}\left(\mathrm{Stroke\ Contrast}_{\mathrm{ij}}\right) \\ & + U_{3\mathrm{j}}\left(\mathrm{Serif\ Type}_{\mathrm{ij}}\right)^{*}\left(\mathrm{Stroke\ Contrast}_{\mathrm{ij}}\right) + R_{\mathrm{ij}}\end{aligned}$$

Let Font-size Threshold$_{ijk}$ denote the kth replicate for the ith participant in the jth group. That is, i = participant level, j = group level, k = population level, U = level-two error, R = population-level error.

2.2. Results

2.2.1. Font-size threshold (α)

In terms of the serif factor, there was no compelling evidence for the difference between the serif and sans-serif font conditions ($E(\mu_{serif} - \mu_{sans}) = 0.98$, 95% Cr. I. = $[-1.29, 3.26]$, $P(\delta > 0) = 0.76$). We concluded that the data and the model did not support H1. The low stroke-contrast font condition produced lower font-size thresholds ($E(\mu_{low}) = 58$ pts., 95% Cr. I. = $[51, 65]$) than the high stroke-contrast font condition ($E(\mu_{high}) = 61$ pts., 95% Cr. I. = $[54, 69]$). There was compelling evidence for this difference ($E(\mu_{low} - \mu_{high}) = -2.92$, 95% Cr. I. = $[-0.64, -5.52]$, $P(\delta > 0) = 0.98$; see Fig. 5), thus, we concluded that the data and the model supported H2. Regarding the interaction between stroke contrast and serif, there was also no compelling evidence that the difference between these conditions is larger than zero ($E(\mu_{low,\ sans} - \mu_{high,\ serif}) = 0.17$, 95% Cr. I. = $[-3, 3]$, $P(\delta > 0) = 0.53$). We concluded that the data and the model did not support H3. For Bayesian pairwise comparisons see Table 1.

2.2.2. Sensitivity (d prime/d′)

The sans-serif font condition produced higher sensitivity ($E(\mu_{sans}) = 1.67$, 95% Cr. I. = $[1.50, 1.84]$) than the serif font condition ($E(\mu_{serif}) = 1.70$, 95% Cr. I. = $[1.53, 1.87]$). There was no compelling evidence for this difference ($E(\mu_{sans} - \mu_{serif}) = -0.03$, 95% Cr. I. = $[-0.08, 0.02]$, $P(\delta > 0) = 0.03$), thus, we concluded that the data and the model did not support H1. The low stroke-contrast font condition produced higher sensitivity ($E(\mu_{low}) = 1.70$, 95% Cr. I. = $[1.53, 1.87]$) than the high stroke-contrast font condition ($E(\mu_{high}) = 1.67$, 95% Cr. I. = $[1.50, 1.84]$). However, there was no compelling evidence for this difference ($E(\mu_{low} - \mu_{high}) = 0.03$, 95% Cr. I. = $[-0.08, 0.02]$, $P(\delta > 0) = 0.02$), thus, we concluded that the data and the model did not support H2.

In terms of the interaction between stroke contrast and serif, the sans-serif font condition with low stroke contrast yielded the highest sensitivity ($E(\mu_{low,\ sans}) = 1.76$, 95% Cr. I. = $[1.58, 1.90]$), followed by the serif font condition with high stroke contrast ($E(\mu_{high,\ serif}) = 1.71$, 95% Cr. I. = $[1.54, 1.89]$), then the sans-serif font condition with high stroke contrast yielded a mean of ($E(\mu_{high,\ sans}) = 1.65$, 95% Cr. I. = $[1.47, 1.82]$), finally, the serif condition with low stroke contrast resulted in a mean of ($E(\mu_{low,\ serif}) = 1.63$, 95% Cr. I. = $[1.46, 1.81]$; see Fig. 6).

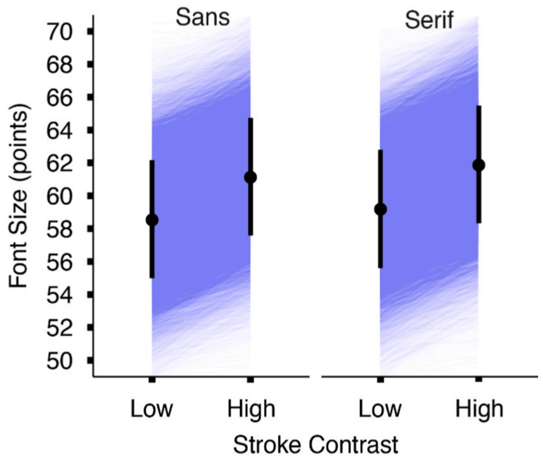


Fig. 5. Font-size threshold as a function of serif type and stroke contrast. Vertical Bars represent the 95% Credible Intervals around the estimated marginal means, which are represented by black circles. Blue areas represent the posterior distribution. Note the only significant effect was the main effect of stroke contrast. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Pairwise comparisons for font-size threshold as a function of serif type and stroke contrast. Cr. I. = credible interval; pd = probability of direction; ROPE = region of practical equivalence. Grey rows represent statistically equivalent conditions and italicised font represents non-significant simple effects.

Font-size threshold pairwise comparisons					
Level 1	Level 2	Difference	95% Cr. I.	pd	% in ROPE
Sans, low	Sans, high	−0.61	(−3.13, 2.21)	0.67	4.98
Sans, low	*Serif, low*	*−2.58*	*(−5.36, 0.00)*	*0.97*	*0.67*
Sans, low	Serif, high	−3.27	(−6.22, −0.46)	0.99	0.00
Sans, high	*Serif, low*	*−2.00*	*(−4.84, 0.68)*	*0.92*	*2.35*
Sans, high	*Serif, high*	*−2.70*	*(−5.43, 0.10)*	*0.97*	*1.19*
Serif, low	Serif, high	−0.68	(−3.30, 2.10)	0.69	5.30

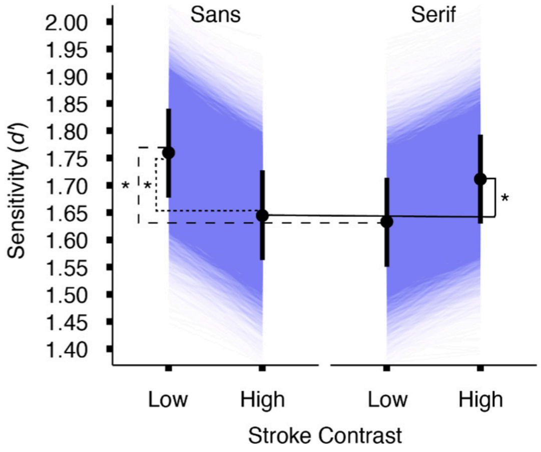


Fig. 6. Sensitivity as a function of serif type and stroke contrast. Asterisks represent significant pairwise comparisons. *$p < .05$. Vertical Bars represent the 95% credible intervals around the estimated marginal means, which are represented by black circles. Blue areas represent the posterior distribution. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

At the level of the sans font condition, the low stroke contrast condition yielded higher sensitivity relative to the high stroke contrast condition ($E(\mu_{sans,\ low} - \mu_{sans,\ high}) = -0.08$, 95% Cr. I. = $[-0.13, -0.02]$, $P(\delta > 0) = 0.99$). At the level of the serif font condition, the opposite pattern was found the low stroke contrast condition yielded lower sensitivity relative to the high stroke contrast condition. However, the difference value did not provide compelling evidence ($E(\mu_{sans,\ low} - \mu_{sans,\ high}) = -0.08$, 95% Cr. I. = $[-0.13, -0.02]$, $P(\delta > 0) = 0.00$). There was compelling evidence for the difference between low stroke contrast and high stroke contrast, when analysed at the level of the sans-serif font condition ($E(\mu_{sans,\ low} - \mu_{sans,\ high}) = -0.13$, 95% Cr. I. = $[-0.18, -0.07]$, $P(\delta > 0) = 0.99$). There was a compelling amount of evidence that the interaction contrast between stroke contrast and serif was larger than zero ($E(\mu_{low,\ sans} - \mu_{high,\ serif}) = 0.19$, 95% Cr. I. = $[0.11, 0.27]$, $P(\delta > 0) = 0.99$). We concluded that the data and the model did support H3.

Table 2
Pairwise comparisons for sensitivity as a function of serif type and stroke contrast. Cr. I. = credible interval; pd = probability of direction; ROPE = region of practical equivalence. Grey rows represent statistically equivalent conditions and italicised font represents non-significant simple effects.

Sensitivity pairwise comparisons					
Level 1	Level 2	Difference	95% Cr. I.	pd	% in ROPE
Sans, low	Sans, high	0.13	(0.06, 0.20)	0.99	21.68
Sans, low	Serif, low	0.11	(0.05, 0.18)	0.99	32.48
Sans, low	Serif, high	0.05	(−0.02, 0.12)	0.91	95.23
Sans, high	Serif, low	−0.01	(−0.09, 0.05)	0.63	99.90
Sans, high	Serif, high	−0.08	(−0.14, −0.01)	0.99	74.39
Serif, low	*Serif, high*	*−0.07*	*(−0.13, 0.01)*	*0.97*	*85.69*

For Bayesian pairwise comparisons see Table 2.

3. Experiment 2

To link our results with the existing LD literature and see if we could replicate the findings of the font-size threshold measure, we ran a classic LD task in Experiment 2 and used identical font stimuli. Based on the size threshold results of Experiment 1, we expected that there will be no evidence for differences between the serif and sans-serif font conditions, low stroke-contrast font conditions will produce lower font-size thresholds, that there will be no interaction between stroke contrast and serif, that pseudowords will yield slower RTs than words. We further expected that word type will interact with a typographical independent variable.

3.1. Method

3.1.1. Participants

A total of 24 individuals were recruited via the same website used for the recruitment of participants for studies (forsøgspersoner.dk). Thus, we collected a convenience sample with the same screening criteria: 18–40 year of age and normal or corrected-to-normal vision. There were 11 females (13 males), and the average age was 26 years (range: 18–37). All participants reported to have normal or corrected-to-normal vision and were paid (i.e., 16 USD) for an hour of participation. All participants considered Danish their primary language. The research followed the tenets of the Declaration of Helsinki.

3.1.2. Design

The follow-up lexical decision task experiment manipulated three repeated-measures, independent variables: (1) serif, (2) stroke contrast, and (3) word type. The serif and stroke contrast factors were identical to those of Experiment 1. The word type was either a word or a pseudoword, which varied on a trial-by-trial basis. The reaction time (RT) served as our dependent variable of interest.

3.1.3. Apparatus

The same experiment set-up and equipment was used for the LDT experiment as that of Experiment 1.

3.1.4. Procedure

Participants were greeted upon arrival and were informed about the nature of the experiment and that their participation was completely voluntary. After the participants gave their informed consent, they were asked to get comfortable in their chair and chinrest. Each participant completed three QUEST procedure runs to obtain their font-size thresholds for a word-recognition task (same task used in Experiment 1). The three thresholds were then averaged and served as the mean result was used as the font size for all subsequent stimuli. They, then, completed a LD task whereby a word or pseudoword was presented until the participant classified the stimulus as a word or a pseudoword by pressing the left arrow key or the right arrow key. Participants were instructed to make a response as fast and as accurately as possible. The stimulus-response mapping was reversed for half of the participants to check for counterbalancing issues. After the response was collected, an inter-trial interval was randomly selected from an exponential distribution, which contained a mean of 300 ms with a lower bound of 300 ms and an upper bound of 800 ms. A total of 800 trials were split up into eight equally-sized blocks and the total testing time ranged between 45 and 55 min.

3.1.5. Data analysis

The same Bayesian framework was used as that of Experiment 1 and the following model was estimated for the reaction time dependent variable:

Level 1:

$$\begin{aligned}\text{Reaction Time}_{\text{ijk}} = {} & \beta_{0\text{j}} + \beta_{1\text{j}}(\text{Serif Type}) + \beta_{2\text{j}}(\text{Stroke Contrast}) \\ & + \beta_{3\text{j}}(\text{Word Type}) + \beta_{4\text{j}}(\text{Serif Type})^*(\text{Stroke Contrast}) \\ & + \beta_{\text{j}}(\text{Serif Type})^*(\text{Stroke Contrast}) + R_{\text{ijk}}\end{aligned}$$

Level 2:

$$\beta_{0\text{j}} = \gamma_{00} + U_{0\text{j}}$$

$$\beta_{1\text{j}} = \gamma_{10}(\text{Serif Type}) + U_{1\text{jk}}$$

$$\beta_{2\text{j}} = \gamma_{20}(\text{Stroke Contrast}) + U_{2\text{jk}}$$

$$\beta_{3\text{j}} = \gamma_{30}(\text{Word Stimulus}) + U_{3\text{jk}}$$

$$\beta_{4\text{j}} = \gamma_{40}(\text{Serif Type})^*(\text{Stroke Contrast}) + U_{4\text{jk}}$$

$$\beta_{5\text{j}} = \gamma_{50}(\text{Serif Type})^*(\text{Word Type}) + U_{5\text{jk}}$$

$$\beta_{6\text{j}} = \gamma_{60}(\text{Stroke Contrast})^*(\text{Word Type}) + U_{6\text{jk}}$$

$$\beta_{7\text{j}} = \gamma_{70}(\text{Serif Type})^*(\text{Stroke Contrast})^*(\text{Word Type}) + U_{7\text{jk}}$$

Full Equation:

$$\begin{aligned}\text{Reaction Time}_{\text{ijk}} = {} & \gamma_{00}(\text{Intercept}) + \gamma_{10}\left(\text{Serif Type}_{\text{ij}}\right) + \gamma_{20}\left(\text{Stroke Contrast}_{\text{ij}}\right) + \gamma_{30}\left(\text{Word Stimulus}_{\text{ij}}\right) + \gamma_{40}\left(\text{Serif Type}_{\text{ij}}\right)^*\left(\text{Stroke Contrast}_{\text{ij}}\right) \\ & + \gamma_{50}\left(\text{Serif Type}_{\text{ij}}\right)^*\left(\text{Word Type}_{\text{ij}}\right) + \gamma_{60}\left(\text{Stroke Contrast}_{\text{ij}}\right)^*\left(\text{Word Type}_{\text{ij}}\right) + \gamma_{70}\left(\text{Serif Type}_{\text{ij}}\right)^*\left(\text{Stroke Contrast}_{\text{ij}}\right)^*\left(\text{Word Type}_{\text{ij}}\right) + U_{0\text{j}}(\text{Intercept}) \\ & + U_{1\text{jk}}\left(\text{Serif Type}_{\text{ij}}\right) + U_{2\text{jk}}\left(\text{Stroke Contrast}_{\text{ij}}\right) + U_{3\text{jk}}\left(\text{Word Type}_{\text{ij}}\right) + U_{4\text{jk}}\left(\text{Serif Type}_{\text{ij}}\right)^*\left(\text{Stroke Contrast}_{\text{ij}}\right) + U_{5\text{jk}}\left(\text{Serif Type}_{\text{ij}}\right)^*\left(\text{Word Type}_{\text{ij}}\right) \\ & + U_{6\text{jk}}\left(\text{Stroke Contrast}_{\text{ij}}\right)^*\left(\text{Word Type}_{\text{ij}}\right) + U_{7\text{jk}}\left(\text{Serif Type}_{\text{ij}}\right)^*\left(\text{Stroke Contrast}_{\text{ij}}\right)^*\left(\text{Word Type}_{\text{ij}}\right) + R_{\text{ijk}}\end{aligned}$$

Let Reaction Time$_{ijk}$ denote the kth replicate for the ith participant in the jth group. That is, i = participant level, j = group level, k = population level, U = level-two error, R = population-level error.

3.2. Results

In terms of the serif factor, there was no compelling evidence for the difference between the serif and sans-serif font conditions (E($\mu_{\text{sans}} - \mu_{\text{serif}}$) = −0.007, 95% Cr. I. = [−0.24, 0.01], $P(\delta > 0) = 0.91$). We conclude that the data and the model support the results of the threshold

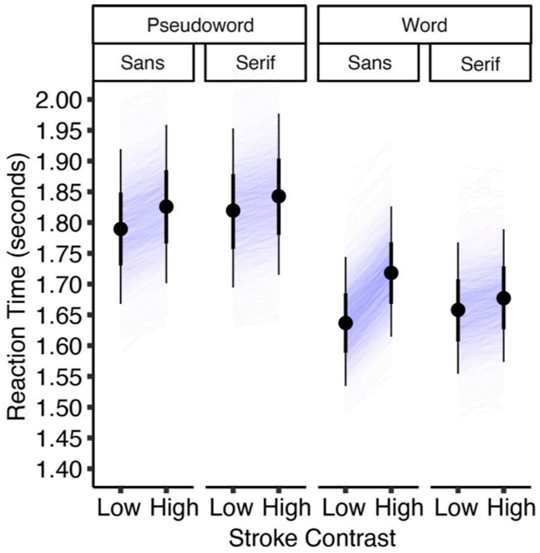


Fig. 7. Reaction time as a function of serif type, stroke contrast, and word type. Vertical Bold and non-bold Bars represent the 95% and 90% Credible Intervals, respectively. The estimated marginal means are represented by the black circles. Blue areas represent the posterior distribution. Note, the significant main effects of stroke contrast (low stroke contrast elicited faster RT than high stroke contrast) and word type (words elicited faster RT than pseudowords), as well as the significant stroke contrast by word type interaction was the main effect of stroke contrast. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

measure of Experiment 1. The low stroke-contrast font condition produced faster RTs (E(μ_{low}) = 1.73 s, 95% Cr. I. = [1.62, 1.84]) than the high stroke-contrast font condition (E(μ_{high}) = 1.77 s, 95% Cr. I. = [1.66, 1.89]). There was compelling evidence for this difference (E($\mu_{\text{low}} - \mu_{\text{high}}$) = −0.43, 95% Cr. I. = [−0.64, −0.22], $P(\delta > 0) = 0.98$), thus, we concluded that the data and the model supported the results of the threshold measure of Experiment 1. There was no compelling evidence, for the interaction between serif and stroke contrast, that the difference between these conditions was larger than zero (E($\mu_{\text{low, sans}} - \mu_{\text{high, serif}}$) = −0.10, 95% Cr. I. = [−0.04, 0.03], $P(\delta > 0) = 0.67$). We concluded that the data and the model supported the results of the threshold measure of Experiment 1. The word condition elicited faster RTs (E(μ_{word}) = 1.68 s, 95% Cr. I. = [1.57, 1.78]) than the pseudoword condition (E($\mu_{\text{pseudoword}}$) = 1.82 s, 95% Cr. I. = [1.69, 1.94]). That is, a significant word-superiority effect was found as expected. There was compelling evidence for this difference (E($\mu_{\text{pseudoword}} - \mu_{\text{word}}$) = 0.148, 95% Cr. I. = [0.09, 0.21], $P(\delta > 0) = 0.99$; see Fig. 7).

Regarding the interaction between serif and word type, there was no compelling evidence that the difference between these conditions was larger than zero (E($\mu_{\text{pseudoword, sans}} - \mu_{\text{word, serif}}$) = 0, 95% Cr. I. = [−0.04, 0.04], $P(\delta > 0) = 0.53$). We concluded that the data and the model followed the expected pattern.

The interaction between word type and stroke contrast yielded compelling evidence that the difference between these conditions was larger than zero (E($\mu_{\text{low, pseudoword}} - \mu_{\text{high, word}}$) = 0.05, 95% Cr. I. = [0.01, 0.09], $P(\delta > 0) = 0.99$). Specifically, the word-superiority effect was larger for the low stroke contrast condition (0.161 s, 95% Cr. I. = [0.10, 0.23]) than for the high stroke contrast condition (0.135 s, 95% Cr. I. = [0.80, 0.20]). We concluded that the data and the model followed the expected pattern. For Bayesian pairwise comparisons see Table 3.

Finally, the three-way interaction between serif, stroke contrast and word type, there was slight evidence that the difference between these conditions was smaller than zero (E($\mu_{\text{low, sans}} - \mu_{\text{high, serif}}$) = −0.05, 95% Cr. I. = [−0.10, 0], $P(\delta < 0) = 0.94$). We concluded that the data and the model followed the expected pattern.

Table 3
Pairwise comparisons for reaction time as a function of serif type, stroke contrast, and word type. Cr. I. = credible interval; pd = probability of direction; ROPE = region of practical equivalence. Bold font represents a significant difference.

Reaction time pairwise comparisons					
Level 1	Level 2	Difference	95% Cr. I.	pd	% in ROPE
Sans, low, pseudo	Serif, low, pseudo	−0.02	(−0.06, 0.01)	0.91	100.00
Sans, low, pseudo	Sans, high, pseudo	−0.03	(−0.07, 0.00)	0.98	100.00
Sans, low, pseudo	Serif, high, pseudo	−0.05	(−0.08, −0.01)	1.00	100.00
Sans, low, pseudo	**Sans, low, word**	**0.16**	**(0.10, 0.23)**	**1.00**	**0.84**
Sans, low, pseudo	Serif, low, word	0.14	(0.07, 0.20)	1.00	9.81
Sans, low, pseudo	Sans, high, word	0.08	(0.02, 0.14)	0.99	78.06
Sans, low, pseudo	Serif, high, word	0.11	(0.05, 0.17)	1.00	36.44
Serif, low, pseudo	Sans, high, pseudo	−0.01	(−0.05, 0.02)	0.76	100.00
Serif, low, pseudo	Serif, high, pseudo	−0.03	(−0.06, 0.01)	0.95	100.00
Serif, low, pseudo	**Sans, low, word**	**0.18**	**(0.11, 0.25)**	**1.00**	**0.00**
Serif, low, pseudo	**Serif, low, word**	**0.16**	**(0.09, 0.23)**	**1.00**	**1.18**
Serif, low, pseudo	Sans, high, word	0.10	(0.04, 0.16)	1.00	48.15
Serif, low, pseudo	Serif, high, word	0.13	(0.08, 0.20)	1.00	10.55
Sans, high, pseudo	Serif, high, pseudo	−0.01	(−0.04, 0.02)	0.82	100.00
Sans, high, pseudo	**Sans, low, word**	**0.20**	**(0.12, 0.27)**	**1.00**	**0.00**
Sans, high, pseudo	**Serif, low, word**	**0.17**	**(0.10, 0.24)**	**1.00**	**0.00**
Sans, high, pseudo	Sans, high, word	0.11	(0.05, 0.18)	1.00	32.99
Sans, high, pseudo	Serif, high, word	0.14	(0.09, 0.21)	1.00	4.50
Serif, high, pseudo	**Sans, low, word**	**0.21**	**(0.14, 0.29)**	**1.00**	**0.00**
Serif, high, pseudo	**Serif, low, word**	**0.19**	**(0.12, 0.26)**	**1.00**	**0.00**
Serif, high, pseudo	Sans, high, word	0.13	(0.06, 0.19)	1.00	18.34
Serif, high, pseudo	**Serif, high, word**	**0.16**	**(0.10, 0.22)**	**1.00**	**0.55**
Sans, low, word	Serif, low, word	−0.02	(−0.06, 0.00)	0.94	100.00
Sans, low, word	Sans, high, word	−0.08	(−0.12, −0.05)	1.00	85.61
Sans, low, word	Serif, high, word	−0.05	(−0.09, −0.01)	1.00	100.00
Serif, low, word	Sans, high, word	−0.06	(−0.09, −0.03)	1.00	100.00
Serif, low, word	Serif, high, word	−0.03	(−0.06, 0.01)	0.94	100.00
Sans, high, word	Serif, high, word	0.03	(0.00, 0.06)	0.98	100.00

4. Discussion

In a word-recognition task and in a classic LD task, we measured reading of four fonts that varied in terms of the presence or absence of serifs and low or high levels of stroke contrast. The font-size threshold measure of both experiments represents the minimum font-size required to get a 75% performance level. Specifically, the lower the font-size threshold value is, for a given font condition, the better the performance is for that font. This measure assumed bias-free performance, however, for Experiment 1 the results can be considered invalid if participants utilized different strategies in each condition or were more inclined to choose one interval over another. To overcome this statistical confound, sensitivity was also calculated and represents performance after potential biases have been accounted for and removed from the analysis.

There was a word-superiority effect in both experiments, which is promising because this effect is expected in LD tasks. The effect in the font-size threshold measures was moderated by the stroke contrast factor, which revealed that low stroke contrast condition yielded better performance when compared to the high stroke contrast condition, and that serifs neither interacted with word type nor stroke contrast. These results were replicated between the two experiments on font-size threshold dependent measure but differed from our sensitivity dependent measure of Experiment 1. This alludes to higher diagnosticity of our 2IFC word-recognition paradigm, which revealed a significant serif by stroke contrast interaction that was not found with the traditional LD task.

Regarding the font-size threshold dependent measure, it was first expected that the sans fonts would yield smaller font-size thresholds relative to the serif fonts (H1); however, there was no evidence for an effect of serifs. It was also expected that the low stroke contrast fonts would elicit lower font-size thresholds, relative to the high stroke contrast fonts, which was supported with a significant main effect of stroke contrast (H2). Thus, words set in fonts with low stroke contrast could be read at smaller font sizes when compared to the words set in fonts with high stroke contrast. This is independent of whether the fonts had serifs or not, although the result was driven by the sans-serif font conditions. It was further expected that the two factors would interact (H3); however, there was no support for this hypothesis in any of the font-size threshold measures.

The data for the sensitivity measure of Experiment 1 differed from that of the font-size threshold dependent measure. The opposite pattern of results from the sensitivity measure was expected because high d-prime values correspond to higher performance and vice versa.

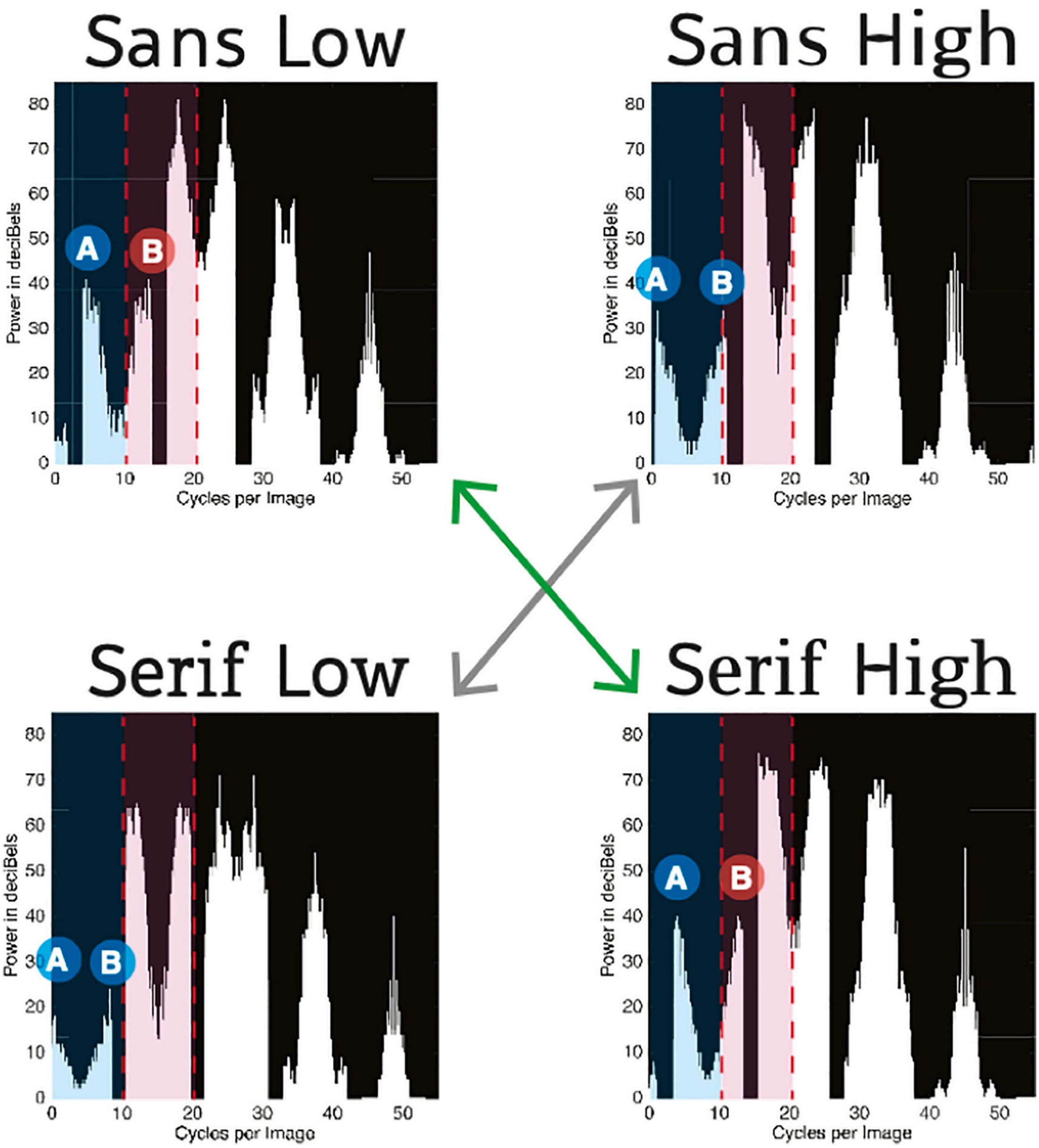


Fig. 8. Power spectral density as a function of frequency (Hz) for the four font condition's alphabets. Note how the power spectrum is similar between sans serif with high stroke contrast and serif with low stroke contrast and between sans serif with low stroke contrast and serif with high stroke contrast. These differences relate to the spatial frequency information provided by the addition of serifs.

Although both independent variables followed the expected pattern (i.e., sans would yield higher d-prime values than serif fonts (H1) and low stroke contrast would yield higher d-prime values than high stroke contrast (H2)), there was no statistical support for these hypotheses.

Our hypothesis that serif style and stroke contrast would interact was supported (H3). That is, stroke contrast moderated the effect of absence of presence of serifs. At the level of the sans-serif font condition, low stroke contrast improved sensitivity when compared to the high stroke contrast condition (partly supporting H2). The pattern was reversed at the level of the serif font condition, however. High stroke contrast tended towards greater sensitivity when compared to low stroke contrast. If we look at the variables the other way around, stroke contrast drew the effect that high stroke contrast test fonts yield greater sensitivity with serifs, while low stroke contrast test fonts yield greater sensitivity with a sans-serif font.

While others have demonstrated significant effects on word-recognition in favour of sans-serif fonts (Garvey et al., 2016; Moret-Tatay & Perea, 2011; Morris et al., 2002), both our font-size threshold and our sensitivity data showed no evidence for the difference between serif and sans-serif fonts, although our findings did follow the hypothesized data pattern. Previous work based on eye-tracking and reading speed paradigms, also found no effect of serif (Akhmadeeva et al., 2012; Arditi & Cho, 2005b; Perea, 2013).

Previous studies have shown that light-weight sans-serif fonts had a negative impact on single-letter visual acuity (Beier & Oderkerk, 2019) and resulted in greater cognitive load by causing longer fixation durations and lower saccadic amplitude (Burmistrov et al., 2016). The test fonts of these previous experiments had similar stroke weight throughout, while our high stroke-contrast fonts only had thin strokes in parts of the letters. Our findings on the sans-serif fonts suggest that the negative impact of lighter-weight fonts, which others have found on sans-serif fonts, can also exist when only parts of the letters have thin strokes. Our findings are further supported by recent work demonstrating that bold serif fonts with thin hairline strokes result in inferior letter recognition (Beier & Oderkerk, 2021). With this experiment, we showed that the effect can also be found in a word recognition task and when using regular-weight sans-serif fonts.

The most surprising result of our study is the reversed pattern between sans-serif and serif fonts. While many experiments concerning font legibility have either compared different fonts (Bernard et al., 2002; Boyarski et al., 1998; Sheedy et al., 2005) or have isolated one typographic variable for investigation (Akhmadeeva et al., 2012; Beier & Dyson, 2014; Moret-Tatay & Perea, 2011; Morris et al., 2002; Perea, 2013), we were able to isolate, simultaneously, two typographic variables. This allowed us to compare the effect of the variables and investigate any possible interaction between the two factors, which we also found with the d-prime measure. Additionally, it was possible to discern which factor was driving the interaction.

When presenting visual stimuli at small visual angles, our visual system draws on its lower spatial frequency channels, which causes letters to appear blurred (Majaj et al., 2002). It is possible that the spatial frequency information can explain the results. Our font-size threshold measure did not capture the moderating effect that stroke contrast had on serif type. That is to say, the sans serif condition's data pattern was in line with our hypothesis (i.e., low stroke contrast yielded better sensitivity than high stroke contrast) and the serif data pattern yielded the opposite effect when it was moderated by stroke contrast (viz., low stroke contrast provided lower sensitivity values, when compared to high stroke contrast). One possible explanation for our sensitivity measure's results is that the addition of serifs resulted in a medium level of spatial frequency contrast (see Fig. 8). In terms of the serif font with high stroke contrast, it is possible that the addition of serifs to a high stroke-contrast font results in a font with an even higher spatial frequency contrast, when compared to the sans-serif fonts, which are not altered by the high spatial frequency information provided by the serifs. Previous work on the effects of spatial-frequency masking of text found that unmasked text reading speed performance cannot be matched irrespective of whether the mask spatial-frequency is a low or high band filter (Beckmann et al., 1991). This suggests that a reader with normal vision must draw on multiple spatial frequency channels, simultaneously, for optimal reading performance. As our two best performing fonts showed greater distribution across the power spectral density (Fig. 8), this could be an explanation for these two fonts leading to better performance when taking the sensitivity measure into account.

Potential limitations of our experiment concern our serif and stroke contrast manipulations, which can be addressed by follow-up experiments that contain a finer-grained manipulation of serifs and stroke contrast. Our fonts' serifs minimally varied regarding contrast; low stroke contrast serifs had lower contrast than the high stroke contrast font. Another potential limitation is that our dependent measures were collected at threshold level, which is not the best performance because, by definition, a threshold is the minimum amount of stimulus energy needed to achieve 75% performance for a 2IFC task (as in Experiment 1). Future experiments can examine the effects of serif and stroke contrast with stimuli that are presented well-above threshold (i.e., supra-threshold) performance because at-threshold perception can vary drastically from supra-threshold perception. We measured word-recognition and LD at size threshold, which has practical implications in relation to setting text on signage and for small font sizes displayed on paper and screen. As mentioned above, it is likely that the results would differ if the experiment was repeated with larger font sizes (i.e., suprathreshold) and a shorter exposure duration.

5. Conclusion

In two investigations (word-recognition and lexical decision tasks) with a font-size threshold measure we found that fonts of low stroke contrast in general was read at smaller font sizes compared to fonts of high stroke contrast.

With a sensitivity (d-prime) measure we found an interaction between the factors of serif and stroke contrast. Sans-serif fonts were read at smaller font sizes when having low stroke contrast, the data for the serif fonts followed a reverse pattern in being read at smaller sizes when having high stroke contrast. Looking at the variables the other way around, stroke contrast was driving the effect that high stroke contrast fonts were read at smaller font sizes when set in serifs while low stroke contrast was read at smaller font sizes when set in sans serif. We offer the explanation that the results are driven by different font styles drawing on different spatial frequency channels.

Declaration of competing interest

Katsumi Minakata and Sofie Beier declare that they have no conflict of interest.

Acknowledgement

The authors thank Bart Cooreman for support in running Experiment 2.

Statement of relevance

With the development of electronic reading devices, users often have the option of customizing the display with the font style of their liking. Our results support the user in making such informed font choices, and provides professionals working within typography with new tools when designing for small visual angles (e.g., footnote text or traffic signage). The study moves away from an often-seen recommendation of specific fonts for reading, towards a recommendation of font characteristics, instead. This allows for better font choices and a better use of the many available fonts on the market.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.actpsy.2022.103623.

References

Akhmadeeva, L., Tukhvatullin, I., & Veytsman, B. (2012). Do serifs help in comprehension of printed text? An experiment with cyrillic readers. *Vision Research, 65*, 21–24.

Arditi, A., & Cho, J. (2005a). Serifs and font legibility. *Vision Research, 45*(23), 2926–2933.

Arditi, A., & Cho, J. (2005b). Serifs and font legibility. *Vision Research, 45*(23), 2926–2933. https://doi.org/10.1016/j.visres.2005.06.013

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278.

Beckmann, P. J., Legge, G. E., & Luebker, A. (1991). *7.5: Reading: Letters, words, and their spatial-frequency content*.

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

Beier, S. (2017). *Type tricks: Your personal guide to type design*. BIS Publishers.

Beier, S., Bernard, J.-B., & Castet, E. (2018). Numeral legibility and visual complexity. *DRS Design Research Society*. https://doi.org/10.21606/drs.2018.246

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

Beier, S., & Oderkerk, C. A. T. (2019). Smaller visual angles show greater benefit of letter boldness than larger visual angles. *Acta Psychologica, 199*, Article 102904. https://doi.org/10.1016/j.actpsy.2019.102904

Beier, S., & Oderkerk, C. A. T. (2021). High letter stroke contrast impairs letter recognition of bold fonts. *Applied Ergonomics, 97*.

Bernard, J. B., & Chung, S. T. (2011). The dependence of crowding on flanker complexity and target–flanker similarity. *Journal of Vision, 11*(8), 1–16.

Bernard, J.-B., Kumar, G., Junge, J., & Chung, S. T. (2013). The effect of letter-stroke boldness on reading speed in central and peripheral vision. *Vision Research, 84*, 33–42. https://doi.org/10.1016/j.visres.2013.03.005

Bernard, M., Lida, B., Riley, S., Hackler, T., & Janzen, K. (2002). A comparison of popular online fonts: Which size and type is best. *Usability News, 4*(1).

Beymer, D., Russell, D., & Orton, P. (2008). In *An eye tracking study of how font size and type influence online reading* (pp. 15–18).

Boyarski, D., Neuwirth, C., Forlizzi, J., & Regli, S. H. (1998). A study of fonts designed for screen display. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '98* (pp. 87–94).

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*(4), 433–436. https://doi.org/10.1163/156856897x00357

Bürkner, P.-C. (2017). Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80*(1), 1–28.

Burmistrov, I., Zlokazova, T., Ishmuratova, I., & Semenova, M. (2016). In *Legibility of light and ultra-light fonts: Eye tracking study* (pp. 1–6). https://doi.org/10.1145/2971485.2996745

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Grantee Submission, 76*(1), 1–32.

Chung, S. T., & Bernard, J.-B. (2018). Bolder print does not increase reading speed in people with central vision loss. *Vision Research, 153*, 98–104. https://doi.org/10.1016/j.visres.2018.10.012

Dogusoy, B., Cicek, F., & Cagiltay, K. (2016). How serif and sans serif typefaces influence reading on screen: An eye tracking study. In A. Marcus (Ed.), *Design, User Experience, and Usability: Novel User Experiences* (Vol. 9747, pp. 578–586). Springer International Publishing.

Frutiger, A., Besset, M., Ruder, E., & Schneebeli, H. R. (1980). *Type, sign, symbol*. ABC Edition.

Garvey, P. M., Eie, W.-Y., & Klenna, M. J. (2016). The effect of font characteristics on large format display. *Interdisciplinary Journal of Signage and Wayfinding, 1*(1).

Kingdom, F., & Prins, N. (2010). In *Goodness-of-fit. Psychophysics: A practical introduction* (pp. 226–228).

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in Psychtoolbox-3? *Perception, 36*(14), 1–16.

Lund, O. (1999). *Knowledge construction in typography: The case of legibility research and the legibility of sans serif typefaces* [Doctoral dissertation]. University of Reading.

Macaya, M., & Perea, M. (2014). Does bold emphasis facilitate the process of visual-word recognition? *The Spanish Journal of Psychology, 17*.

Macmillan, N., & Creelman, C. (1991). *Detection theory: A user's guide*. Cambridge: Cambridge University Press.

Majaj, N. J., Pelli, D. G., Kurshan, P., & Palomares, M. (2002). The role of spatial frequency channels in letter identification. *Vision Research, 42*(9), 1165–1184.

Makowski, D., Ben-Shachar, M. S., Chen, S., & Lüdecke, D. (2019). Indices of effect existence and significance in the Bayesian framework. *Frontiers in Psychology, 10*, 2767.

Makowski, D., Lüdecke, D., & Ben-Shachar, M. (2020). *Modelbased: Estimation of model-based predictions, contrasts and means*.

McLean, R. (1980). *The Thames and Hudson manual of typography*. Thames and Hudson.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology, 90*(2), 227.

Moret-Tatay, C., & Perea, M. (2011). Do serifs provide an advantage in the recognition of written words? *Journal of Cognitive Psychology, 23*(5), 619–624.

Morris, R. A., Aquilante, K., Yager, D., & Bigelow, C. (2002). In *, 33. P-13: Serifs slow RSVP reading at very small sizes, but don't matter at larger sizes* (pp. 244–247). https://doi.org/10.1889/1.1830242

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*(4), 437–442.

Pelli, D. G., Burns, C. W., Farell, B., & Moore-Page, D. C. (2006). Feature detection and letter identification. *Vision Research, 46*(28), 4646–4674. https://doi.org/10.1163/156856897x00366

Perea, M. (2013). Why does the APA recommend the use of serif fonts? *Psicothema, 25*(1), 13–17.

Prins, N., & Kingdom, F. (2018). Applying the model-comparison approach to test specific research hypotheses in psychophysical research using the Palamedes toolbox. *Frontiers in Psychology, 9*, 1250.

Russell, V. L. (2021). emmeans: Estimated marginal means, aka least-squares means (R package version 1.6.1) [Computer software]. https://CRAN.R-project.org/package=emmeans.

Schütt, H. H., Harmeling, S., Macke, J. H., & Wichmann, F. A. (2016). Painfree and accurate Bayesian estimation of psychometric functions for (potentially) overdispersed data. *Vision Research, 122*, 105–123.

Sheedy, J. E., Subbaram, M. V., Zimmerman, A. B., & Hayes, J. R. (2005). Text legibility and the letter superiority effect. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 47*(4), 797–815. https://doi.org/10.1518/001872005775570998

Sperling, G. (1965). Temporal and spatial visual masking. I. Masking by impulse flashes. *JOSA, 55*(5), 541–559.

Stan Development Team. (2017). Stan: A C++ Library for Probability and Sampling (Version 2.14.0) [Computer software]. http://mc-stan.org/.

Stan Modeling Language. (2017). User's guide and reference manual. http://mc-stan.org/manual.html.

Unger, G. (2007). *While you're reading*. Mark Batty Publisher.

Watson, A. B. (2017). QUEST+: A general multidimensional bayesian adaptive psychometric method. *Journal of Vision, 17*(3), 10. https://doi.org/10.1167/17.3.10

Watson, A. B., & Pelli, D. G. (1983). QUEST: A Bayesian adaptive psychometric method. *Perception & Psychophysics, 33*(2), 113–120.