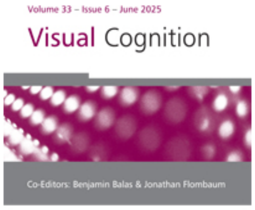
A method for studying the contextual similarity of characters in Cyrillic, Devanagari, and Latin scripts and exploration of the effects of typeface design and expertise

Routledge
Taylor & Francis Group


Volume 33 – Issue 6 – June 2025
Visual Cognition
Co-Editors: Benjamin Balas & Jonathan Flombaum

Visual Cognition

ISSN: 1350-6285 (Print) 1464-0716 (Online) Journal homepage: www.tandfonline.com/journals/pvis20

A method for studying the contextual similarity of characters in Cyrillic, Devanagari, and Latin scripts and exploration of the effects of typeface design and expertise

David Březina

To cite this article: David Březina (16 Feb 2026): A method for studying the contextual similarity of characters in Cyrillic, Devanagari, and Latin scripts and exploration of the effects of typeface design and expertise, Visual Cognition, DOI: 10.1080/13506285.2026.2627306

To link to this article: https://doi.org/10.1080/13506285.2026.2627306

Published online: 16 Feb 2026.

VISUAL COGNITION
https://doi.org/10.1080/13506285.2026.2627306

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

A method for studying the contextual similarity of characters in Cyrillic, Devanagari, and Latin scripts and exploration of the effects of typeface design and expertise

David Březina [a,b*]

[a]Department of Typography & Graphic Communication, University of Reading, Reading, UK; [b]Rosetta Research, Brno, Czech Republic

ABSTRACT

Although perceptual similarity is highly contextual, visual similarity of characters (letters) is typically investigated using stimuli that do not control for contextual effects, e.g., comparison of pairs, identification of single characters. Moreover, most character similarity studies focus on a single typeface (font) or study under conditions that are detrimental to the quality of the designs (e.g., low resolution or low contrast), which makes it challenging to generalize their findings. To obtain data that permit a more detailed and realistic enquiry of character similarity relationships, the study reported in this paper used a contextual similarity task. Participants were presented with character triplets and asked to pick the odd one out, thus judging the remaining characters as more similar. To demonstrate the method's robustness and transferability across scripts, this cognitively simple yet challenging task was used with a diverse selection of typefaces, world scripts, and participants ($n = 1721$). The results showed that the contextual similarity task is sensitive to the effects of typeface design and elicits similarity judgements that are hard to predict using pairwise data. Comparisons across groups of participants showed effects of design expertise, nativity, and fluency in relevant scripts.

ARTICLE HISTORY
Received 21 October 2024
Accepted 1 February 2026

KEYWORDS
Visual similarity; character perception; context effects; odd-one-out task; typeface design

Motivations

Perceptual similarity judgements are contextual by nature (Tversky, 1977). For example, the similarity of a pair “y, z” will likely be higher when judged within the triplet “o, y, z” (due to “o” being rounded unlike “y, z”) and lower within the triplet “s, y, z” (due to the strong similarity of “s” and “z”). Typefaces, collections of geometric descriptions of character shapes, are typically designed with the aim of achieving a unifying sense of style that can be perceived by readers (compare Times New Roman and Comic Sans in Figure 1). Distinct character shapes are designed to be visually coherent with other character shapes from the same typeface (e.g., Baudin, 1989, p. 23; Hofstadter & McGraw, 1998, p. 417).[1] This visual coherence can be seen as a network of contextual similarity relationships among character shapes.

However, the typeface design field is not merely concerned with how typefaces look. Among other things, it is also concerned with the effects particular designs can have on ease of reading. In typographic practice, the reading experience is typically assessed subjectively, with respect to various craft theories, e.g., theories regarding the positive effect of familiarity and adherence to norms (Gill, 1931/1988, p. 44), clearly articulated design features (Unger, 2018, p. 177), or balanced inter-letter spacing (Tracy, 2003, p. 70). Scientists have devised methods to measure various aspects of the reading process (e.g., Beier & Larson, 2013; Dyson, 2019; Grainger et al., 2016). However, it is not always clear how to apply the research results in typeface design practice.

There is strong evidence for recognition/identification of individual characters (rather than word shapes) as the basis for word recognition and reading (e.g., McClelland & Rumelhart, 1981; Pelli et al., 2003; Pelli et al., 2007; White et al., 2008). Successful word recognition appears to be facilitated by low-

CONTACT David Březina ✉ david@mrbrezina.com

*This research was part of the author's work towards the degree of Doctor of Philosophy at the Department of Typography & Graphic Communication, University of Reading (UK). The author is currently affiliated with Rosetta Research.

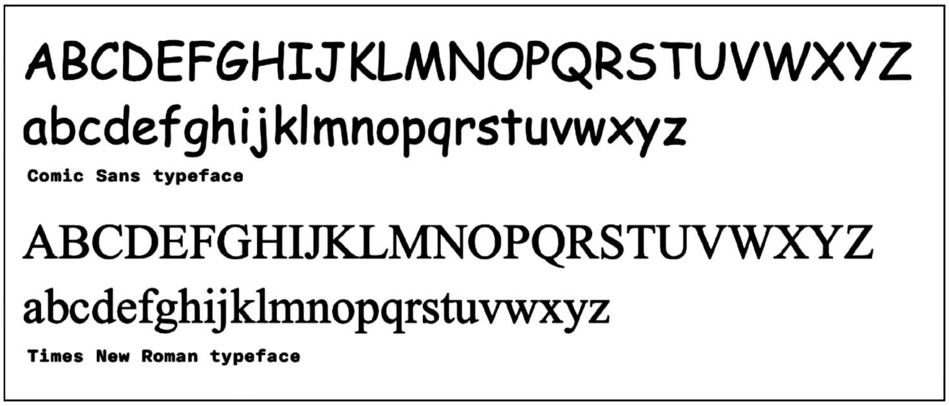

ABCDEFGHIJKLMNOPQRSTUVWXYZ
abcdefghijklmnopqrstuvwxyz
Comic Sans typeface
ABCDEFGHIJKLMNOPQRSTUVWXYZ
abcdefghijklmnopqrstuvwxyz
Times New Roman typeface

Figure 1. Characters from Comic Sans and Times New Roman typefaces have a different visual style.

level visual character recognition and high-level orthographic knowledge that helps clarify character identities in case they are uncertain (Lally & Rastle, 2023; Marcet & Perea, 2017); character shapes that are too similar and can be easily confused make the reading process more difficult. However, it is not clear whether the opposite is also true to some extent, that the right kind of visual coherence among letters can be beneficial to legibility by aiding recognition of otherwise distinct character shapes. This idea was previously explored in font-tuning studies by Sanocki (1987) and reviewed by Walker (2008). It is also supported by typeface design practice where parts of character shapes are reused across different character shapes, e.g., the bottom serifs, vertical stems, and arches in letters "h, m, n" or rounded bowls in "b, c, d, e, o, p, q" in the Times New Roman typeface (see Figure 1). The use of a limited set of building blocks may hypothetically aid character recognition.

Analyzing internal representations through similarity studies with respect to typeface design can thus help explore the relationship between similarity and recognition which would be beneficial to reading research including reading acquisition and education. For example, Pick (1965) shows that distinctive features like crossing lines or loops are used to organize newly learned letter-like shapes by early readers. Moreover, detailed analysis of similarity based on readers' perception will provide grounds for the critique of design decisions in typefaces, potentially connecting results of reading research to actionable design instructions through further testing of typeface design effects on reading. This can feed into the craft's methodology, quality improvements (more legible fonts), design for unfamiliar scripts, or automation of parts of the design process.

The visual domain of all character shapes is diverse in two dimensions: across categories ("a" is different from "b") and within categories ("a" in a serif typeface differs from "a" in a sans-serif typeface) which makes the search for internal representations particularly challenging. Different typefaces and handwriting styles mean that characters are visualized in diverse ways. Hofstadter and McGraw (1998) illustrate this richness by showing diverse shapes of the letter "a" (see Figure 2).

Studies that take a single font and treat it as a sufficient singular representative of a script run the

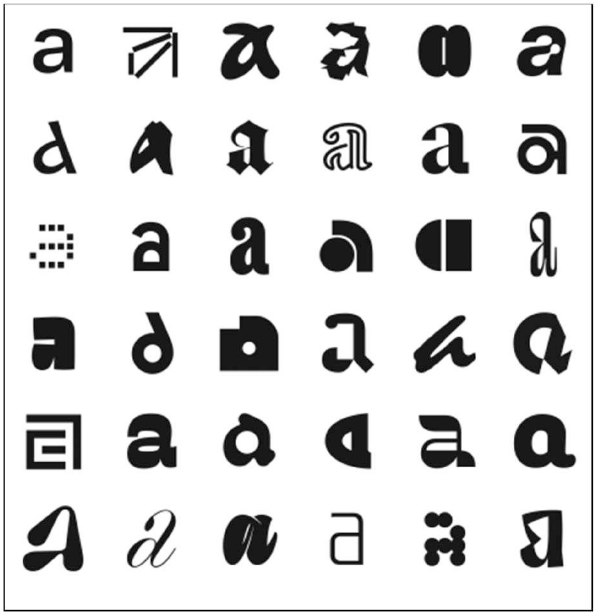


Figure 2. Thirty six different character shapes for the letter "a" taken from contemporary fonts (after an illustration from Hofstadter & McGraw, 1998, p. 413).

risk of oversimplifying the visual domain (e.g., Chang et al., 2018; Changizi et al., 2006; Changizi & Shimojo, 2005; Hinton et al., 1992) by assuming that individual characters have an invariant set of building blocks. Figure 2, a small subset of all the creative possibilities, shows how problematic this assumption is. See also a review of Changizi and Shimojo (2005) by Březina (2021).

It is worth stating that if the character shapes can change significantly within categories (i.e., across typefaces or handwriting styles), the structure of similarity relationships among character shapes within typefaces can change as well.[2] The fact that most of the existing similarity studies conducted over the last hundred years (summarized by Mueller & Weidemann, 2012) use stimuli in a single typeface prevents indiscriminate application of their findings in typeface design or character recognition research.

The ability to recognize characters is learned (Dehaene, 2010; Gibson et al., 1962). The influence of familiarity on character recognition has been documented by Gauthier et al. (2006) who found differences in how experts and novice readers process Chinese character shapes and Wiley et al. (2016) and Wiley and Rapp (2019) who found an effect of familiarity with the Arabic script on participants' perception. Moreover, designers that closely engage with character shapes (typeface designers, lettering artists, typographers) may have a different perception as well and use different assessment strategies (Dyson et al., 2016; Dyson & Stott, 2012). Hence, there may be an influence of language/script or design expertise on the way people make similarity judgements.

With respect to world scripts, most of the studies reported by Mueller and Weidemann (2012) are focused on the Latin script with a preference for uppercase over lowercase letters of the English alphabet.[3] Firstly, focusing on uppercase is not representative of usual Latin long-form texts which consist mostly of lowercase letters (see also Sanocki & Dyson, 2012). Secondly, in a globalized, multilingual world, it is worth asking what linguistic scope similarity studies have and whether the methods and outcomes are specific to individual languages, scripts, or settings and miss consideration of the ways other languages or scripts are organized or used. A few examples that might affect transferability are: the distinction between uppercase and lowercase letters which is relatively rare among world's scripts; use of diacritics (Perea et al., 2021; Simpson et al., 2013); different levels of visual complexity (Chang et al., 2018); presence or lack of mirrored letters (e.g., Tamil, see Blasi et al., 2022), connecting characters, or frequent use of allographs (e.g., Arabic, see Boudelaa et al., 2020).

Terminology

Writing has been studied in linguistics, psychology, computer science, and visual communication with each field concerned only with some of its aspects. Typography and typeface design are concerned with stylistic, technological, and material notions of writing which are not always communicated clearly in linguistics or psychology. Therefore, this paper uses the following terms (from the most general to the most specific):

- *character* refers to a general category, the basic unit of a script, e.g., any letter "a" or any letter "b",
- *character shape* refers to the specific shape of a character, e.g., letter "a" from the Georgia typeface,
- *character image* refers to a specific material rendering of a character shape, e.g., letter "a" as it is rendered in this paper, viewed either printed or on screen.

The use of the term *character* is consistent with its use in the Unicode specification (The Unicode Consortium, 2023; Whistler & Freytag, 2022), *basic shape* in linguistics (Meletis, 2020), or *abstract letter identity* (also symbolic letter identity) in psychology (Rothlein & Rapp, 2017). *Character image* corresponds to the term *graph* as used by Meletis (2020). A *character shape* can correspond to different *character images* depending on production, material, and other visual and environmental factors (for design-technological considerations see Southall, 1986, 2005).[4]

Existing methodologies

The motivations of character-similarity and character-recognition studies range from finding ways to improve legibility and reading education, to more general attempts to explore the cognitive system and internal character representations.

Visual similarity has been examined in studies using character pairs as stimuli. The similarity measures are derived from participants' subjective rating of similarity (Boles & Clifford, 1989; Boudelaa et al., 2020; Kaiho, 1970; Kuennapas & Janson, 1969; Simpson et al., 2013; Wiley et al., 2016) or response times for binary-choice answers or go/no-go task (Courrieu et al., 2004; Higuchi & Kobayashi, 2022; Podgorny & Garner, 1979).

Character pair ratings are necessarily relative to the context provided by all characters studied (see also Wiley & Rapp, 2019). If the complete set of characters is not shown, participants need to rely on their memory for internal representations of conventional character shapes which may be unreliable. Some studies address this issue and make the complete set of stimuli available for the participants to refer back to for comparison (e.g., Kaiho, 1970; Simpson et al., 2013). It is unclear whether participants do so. Either way, participants have to estimate all the relative similarities to other, possibly yet unseen, characters and resolve an implied, complex network of similarity relationships in order to respond with a single relative pair rating. It does not seem safe to assume that even competent readers have perfect knowledge and recollection of all character shapes from a particular typeface (Wong et al., 2018).

Theoretically, similarity measures can also be derived indirectly from errors, confidence ratings, and response times in recognition tasks with individual characters. However, when presenting characters in clear and readable conditions with no time limit for the recognition task, only a few errors are generated which prevents making useful conclusions about character similarity (Simpson et al., 2013). In order to challenge participants' perception, and to therefore produce more errors, researchers use brief presentation (Bouma, 1971; Gilmore et al., 1979; Mueller & Weidemann, 2012; Townsend, 1971), reducing the letter size or increasing reading distance (Bouma, 1971; Phillips et al., 1983), presenting the characters in the peripheral visual field (Alexeeva, 2024; Reich & Bedell, 2000), or making the viewing conditions worse by means of lower contrast (Geyer, 1977) or tight inter-character spacing (Liu & Arditi, 2001). Note, that like pair judgements, tasks with individual characters rely on participants' memory of internal representations.

With the exception of more recent studies (e.g., Boudelaa et al., 2020; Mueller & Weidemann, 2012; Simpson et al., 2013), presentation of stimuli rarely meets the quality standards of contemporary digital fonts: the screen resolutions are low, typefaces have heavily simplified character shapes, overly challenging conditions are used. While acceptable when exploring more general, conceptual relationships, poor presentation may prevent careful assessment of the shapes. Fiset et al. (2009, p. 24) note that, in particular, low contrast and brief presentation "exacerbate the relative importance of low spatial frequencies".

Approach

The feature-based approach formalized by Tversky (1977) is well established within character recognition research (e.g., Biederman, 1987; Fiset et al., 2008; Grainger et al., 2008; Lanthier et al., 2009; Palmer, 1999; Pelli et al., 2006; Petit & Grainger, 2002; Rosa et al., 2016; Wiley et al., 2016). However, it is likely that the features used for recognition differ somewhat from the similarity features due to the differing objectives of the two tasks. Mueller and Weidemann (2012) distinguish between *perceivability* (or legibility) of a character: "a theoretical construct affecting the probability the observer forms a veridical percept from the stimulus, independent of response factors" and its *similarity*: "the distinctiveness of a stimulus within a set of other stimuli". While other researchers derive similarity from perceivability (see character recognition studies above), in their paper, Mueller and Weidemann show that models treating similarity and perceivability as independent factors are more parsimonious and recommend treating these concepts as independent. This makes a compelling case for studies of similarity that do not use a recognition task that could confound observations regarding similarity.

The purpose of this study was to obtain data about perception with presentation sensitive to typeface design effects. Additionally, the need for character identification had to be eliminated to allow readers who might not be familiar with a particular script to participate. These goals were achieved by introducing a contextual similarity task.

The contextual similarity task studies pair similarity within as simple a context as possible: the participants were presented with a group of three characters and tasked to identify the most different character shape from the other two (the odd one out). Assuming similarity judgements are inverse to difference judgements (discussed by Tversky, 1977), this is the same as asking them to select the two most similar characters from the triplet. However, the explanation of the task is much simpler as it builds on participants' ability to discriminate the odd one out learned from previous experience with widely used studies and games using a similar approach. Working with triplets also removes complexities associated with relative ratings and simplifies responses to a straightforward ternary choice.

Studies with a ternary choice are a specific case of a best-worst task (Hollis & Westbury, 2018) and have been used to explore personal constructs (Kelly, 1955), text documents (Schultz & Joachims, 2003), bitmap images (Wang et al., 2016), illustrations (Garces et al., 2014), 3D-object models (Lun et al., 2015), and multi-modal similarity (McFee & Lanckriet, 2011). Tasks with character triplets have not previously been used to study character similarity.

Method

Participants

The experiment was piloted with six participants and using 23 triplets of characters from a single typeface. The participants tended to agree on the more obvious triplet trials and did not have a problem understanding the task and using the website.

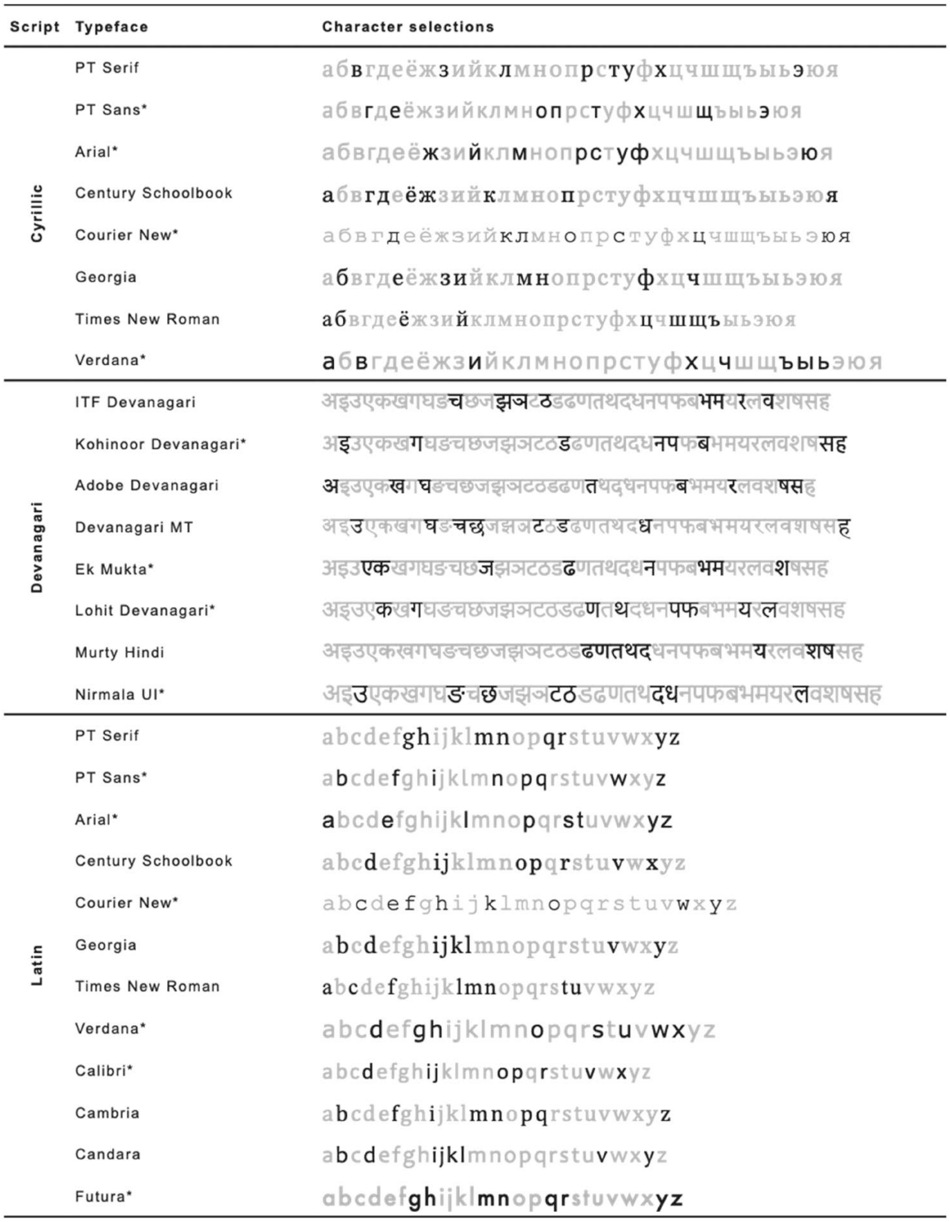
The participants were invited by convenience sampling, social networks, and via websites dedicated to recruiting participants.[5] The only constraint was that the participants needed to be adults which was clearly stated in the introduction to the experiment. Paid advertising was used to target the call for participants to the right audience for the Cyrillic and Devanagari variants, i.e., to people who can speak the relevant languages or live in corresponding countries for each script. There was no financial incentive for taking part. Participants were not prevented from taking part in variants for a script they cannot read, however this was not encouraged. Participants' data was saved only when they had completed the experiment.

Stimuli

The aim of the study was to challenge the experimental method to see whether it can elicit typeface design and expertise effects and work with different scripts. Testing diverse stimuli was more important than providing a complete similarity matrix for a specific language alphabet or script. Twenty eight typefaces intended for continuous reading in Cyrillic, Devanagari, and Latin were used in twenty eight experiment variants (see Figure 3 and appendices for details). The typefaces studied were selected from the pool of typefaces commonly available in operating systems. For the sake of design diversity, half of the typefaces were low-contrast designs and the other half were high-contrast designs (see Figure 3 for explanation of the typographic contrast) and for Latin and Cyrillic, sans serif and serif typefaces were included. Only regular (or regular-like) styles were used.[6] Italic and bold styles were not studied.

As mentioned earlier, similarity and character recognition studies tend to focus on a very limited selection of characters (most often A – Z in English). A study focused on all characters from a specific language alphabet or script would yield too many triplet combinations (e.g., triplet combinations from 26 characters would amount to 2600 triplet trials) which would be extremely time-consuming and tedious for participants to complete, particularly across multiple typefaces. Thus, the selection was limited to eight characters from a particular typeface which gives a series of 56 triplets when exhaustively combined with no character repetition within a triplet. Each series contained only characters from a single typeface and script. Based on the pilot testing, participants can complete a series of 56 triplets in approximately five minutes. Short completion time is beneficial in an online experiment as it prevents fatigue and reduces dropout rates.

The eight characters for each script-typeface combination were selected systematically with preference for lowercase in Latin and Cyrillic and basic consonants in Devanagari (see appendices for details). The goal was to represent characters as evenly as possible across all experiment variants and to combine highly similar characters as well as highly dissimilar characters in each series. The risk of bias was reduced by using a unique selection of characters for most typefaces and combining the characters from the selection exhaustively to produce visually diverse triplet combinations.



Script	Typeface	Character selections
Cyrillic	PT Serif	абвгдеёжзийклмнопрстуфхцчшщъыьэюя
	PT Sans*	абвгдеёжзийклмнопрстуфхцчшщъыьэюя
	Arial*	абвгдеёжзийклмнопрстуфхцчшщъыьэюя
	Century Schoolbook	абвгдеёжзийклмнопрстуфхцчшщъыьэюя
	Courier New*	абвгдеёжзийклмнопрстуфхцчшщъыьэюя
	Georgia	абвгдеёжзийклмнопрстуфхцчшщъыьэюя
	Times New Roman	абвгдеёжзийклмнопрстуфхцчшщъыьэюя
	Verdana*	абвгдеёжзийклмнопрстуфхцчшщъыьэюя
Devanagari	ITF Devanagari	अइउएकखगघङचछजझञटठडढणतथदधनपफबभमयरलवशषसह
	Kohinoor Devanagari*	अइउएकखगघङचछजझञटठडढणतथदधनपफबभमयरलवशषसह
	Adobe Devanagari	अइउएकखगघङचछजझञटठडढणतथदधनपफबभमयरलवशषसह
	Devanagari MT	अइउएकखगघङचछजझञटठडढणतथदधनपफबभमयरलवशषसह
	Ek Mukta*	अइउएकखगघङचछजझञटठडढणतथदधनपफबभमयरलवशषसह
	Lohit Devanagari*	अइउएकखगघङचछजझञटठडढणतथदधनपफबभमयरलवशषसह
	Murty Hindi	अइउएकखगघङचछजझञटठडढणतथदधनपफबभमयरलवशषसह
	Nirmala UI*	अइउएकखगघङचछजझञटठडढणतथदधनपफबभमयरलवशषसह
Latin	PT Serif	abcdefghijklmnopqrstuvwxyz
	PT Sans*	abcdefghijklmnopqrstuvwxyz
	Arial*	abcdefghijklmnopqrstuvwxyz
	Century Schoolbook	abcdefghijklmnopqrstuvwxyz
	Courier New*	abcdefghijklmnopqrstuvwxyz
	Georgia	abcdefghijklmnopqrstuvwxyz
	Times New Roman	abcdefghijklmnopqrstuvwxyz
	Verdana*	abcdefghijklmnopqrstuvwxyz
	Calibri*	abcdefghijklmnopqrstuvwxyz
	Cambria	abcdefghijklmnopqrstuvwxyz
	Candara	abcdefghijklmnopqrstuvwxyz
	Futura*	abcdefghijklmnopqrstuvwxyz

Figure 3. World scripts, typefaces, and character selections. Character selections from the basic alphabets and syllabary that were used in the series of triplet trials for a particular experiment variant are marked in black. Low-contrast typefaces are marked with an asterisk. The remaining typefaces are high-contrast typefaces. The term contrast in typography refers to differentiation between thick and thin strokes of the characters.

Conducting the experiment online meant that there was limited control over viewing distance and environment in which the participants completed the tasks. It is possible that the samples, presented on a website, might have been displayed smaller or larger depending on a participant's device – especially if it was a mobile device – and any custom screen or browser settings. Particular care was given to minimize the effects of perceivability by keeping the viewing conditions comfortable with no additional challenges such as small size, filtering or effects of rasterization on character shapes, or brief exposure.

The sample squares were defined as 250 px tall in the HTML/CSS language which corresponds to 6.6 cm (2.6 in.) on computer screens with a resolution of 96 dpi. This ensured that the main differences between typefaces could be examined reliably across various computing devices. Although less controlled, the results can be considered representative of common viewing conditions and congruent with participants' daily experience.

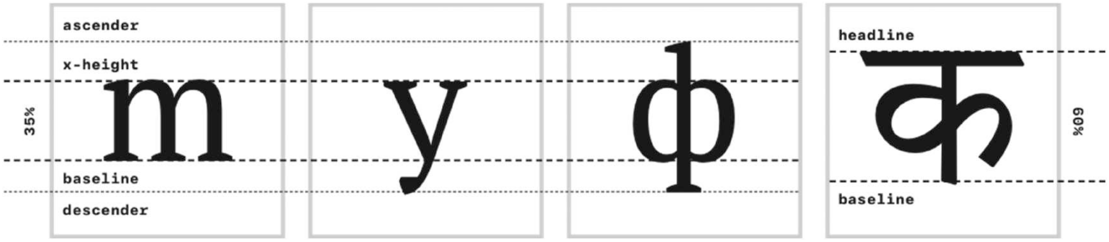


Figure 4. In order to achieve comparable apparent importance across scripts and typefaces, the Latin and Cyrillic characters (first three squares) were scaled so their base height occupied around 35% of the vertical space of the sample square. Balancing base height is a technique recommended for comparison of different typefaces by Smeijers (1996, pp. 33–39). The Devanagari characters (last square) were scaled larger, their base height occupied around 60% of the vertical space. This accounts for their higher visual complexity and larger relative height across a set of fonts that include both, Latin and Devanagari scripts. In the Cyrillic and Latin, the base height refers to the distance between the baseline and the x-height while in the Devanagari, it refers to the distance between the baseline and the *headline* (also called *shirorekha* in Hindi), i.e., the horizontal line connecting Devanagari characters in the top part.

Additionally, characters across all three scripts were presented in a comparable setting respecting their customary sizing in texts for continuous reading (see Figure 4).

Procedure

The experiment was conducted online using a custom-built website titled "letter resemblance study" to collect a large number of responses from participants with diverse professional and linguistic backgrounds and varying degrees of fluency in the scripts studied. There were two parts: a questionnaire and series of triplet trials. The instructions were provided in English. The experiments with Cyrillic and Devanagari stimuli were also made available in Russian and Hindi, respectively.

In order to record potential influences of their background, participants completed a questionnaire to self-report: their nativity and reading fluency with respect to a particular language (and thus a script), their age (as a range), their regular experience as readers, and their experience as designers, if any.

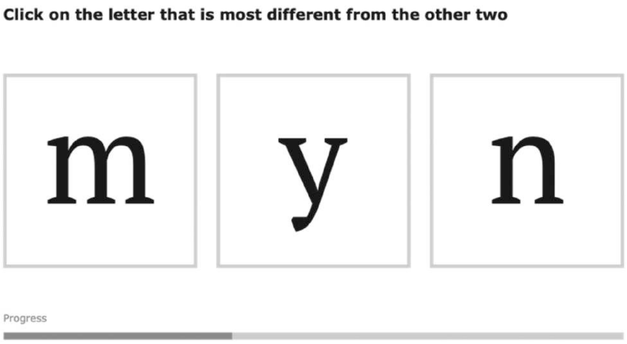

Click on the letter that is most different from the other two

Figure 5. The presentation of triplets during the experiment.

In the second part of the experiment, participants worked through a series of character-triplet trials. The series was introduced by a short instruction page which discouraged participants from being overly fussy about the shape details. Participants were instructed to proceed quickly and consider only the shapes, not the meaning of the characters or character groups. Additionally, it was emphasized that there are no right or wrong answers.

An example of a typical triplet trial is shown in Figure 5. The three characters were presented in three equally sized sample squares to discourage participants from reading the triplet as a single word. The participants were asked to "click on the letter that is most different from the other two". Clicking on the sample square recorded the answer and showed the next triplet from the series.

The triplets were shown to each participant in a different random order. The order of characters within a triplet was randomized as well. There was no option to return back to a previously judged triplet. The option to give no response was not provided to ensure results were obtained even for the more challenging triplets.

Preprocessing

An early form of the experiment included two sessions, one with 56 triplets in a low-contrast typeface, and one with 56 triplets in a high-contrast typeface, both in the same script. This variant of the experiment was eventually abandoned in favour of a single session as this was easier and thus more likely to be completed. Only the first sessions from the two-session variant were used in the following analysis.

Data for a participant that self-reported an empty list of fluent languages were removed.

Statistical analyses

Responses to particular triplet trials are aggregated across participants to obtain response counts (denoted as *c*) and normalized frequencies (denoted as *f*). For each trial, the character with the highest count/frequency is the most popular response or the *overall odd one out* (OOOO).

A trial for a triplet "a, b, c" set in a typeface T where participants agree completely that "c" was the OOOO has frequencies $f(\text{a, b, c})_T = (0.0, 0.0, 1.0)$. On the other hand, a trial with the lowest agreement has frequencies $f(\text{a, b, c})_T = (.33, .33, .33)$.[7] The frequencies for the OOOO in these two extreme cases correspond to the theoretical maximum (1.0) and minimum (.33) of participants' response frequencies, respectively, and they are indicative of how challenging a particular trial is. Participants are not expected to achieve perfect agreement. There are multiple, equally valid, criteria that can be used to judge character similarity. When participants' responses are less reliable, this indicates that trial judgements are more difficult.

When reporting responses per trial, the binomial 95% confidence intervals are provided to quantify the uncertainty of the data with respect to the general population. These are calculated using the Clopper–Pearson interval method (Clopper & Pearson, 1934). The method is exact, not an approximation. It provides satisfactory confidence intervals for the OOOO and works with smaller numbers of participants. The method is relatively conservative compared to other methods.

In order to ascertain the effects different typeface designs may have with otherwise identical triplets, Fisher's (1922) exact test of independence (FET) is used to compare response counts. The counts are scaled to achieve matching totals, an essential requirement of FET.

In order to enquire whether design expertise or script fluence/nativity may have an effect, participants are split into groups according to these criteria. The group responses are investigated with respect to the inter-participant agreement within each group and agreement between aggregated response frequencies for each group. Participants were not responding to identical sets of trials, and the distribution of their response frequencies violated the normality requirement. These conditions make common analyses (ANOVA, covariates) challenging and determined the choice of statistics used.

The inter-participant agreement regarding the complete profile of trial responses, including the OOOO and also the second and third characters, is assessed using Gwet's (2008) agreement coefficient AC1. AC1 accounts for agreement happening by chance and spans values from 0.0 (absence of reliability) to 1.0 (perfect reliability), with *p* indicating the significance of the result. Rather than using AC1 as a measure of the experiment's reliability, it can be taken as an indication of how challenging a particular set of trials is.

Comparisons between groups are reported as percentage agreement regarding the OOOO, i.e., the percentage of shared trials where the OOOO is the same character in both groups. The Mann–Whitney *U* (MWU) test with the null hypothesis of equal distributions is used to compare the OOOO frequencies for each group. However, such a comparison is meaningful only for trials that achieved agreement regarding the OOOO, i.e., the OOOO character being identical in both groups. Other OOOO frequencies are disregarded. In addition to the MWU's *p* value, the associated Common Language Effect Size (CLES) is reported to describe the proportion of frequencies from the first group that is larger than the values of the second group. To complement the group comparisons, Spearman's rank correlation for the complete set of their aggregated response frequencies is also reported.

In order to maintain the reliability of the results, the statistics are reported only when there are at least ten triplets ($m \geq 10$) overlapping between the groups and when each group contains at least five participants ($c \geq 5$). The significance level is set to $\alpha = .05$ for all tests and corrected where necessary.

Results

Demographic data

Table 1 and Figure 6 show participants' responses to the questionnaire. The participants came from diverse linguistic backgrounds. Overall, they self-reported 79 languages to be their native and 88 languages that they could read fluently. There is a

Table 1. The five most common languages in which participants reported as being native/fluent for each script.

Age	Cyrillic	Devanagari	Latin
What is your native language?	Russian (426), Ukrainian (83), Belarusian (27), English (20), Bulgarian (6), other (33)	Hindi (178), English (62), Marathi (56), Tamil (31), Nepali (22), other (156)	English (322), German (122), Dutch (79), French (55), Czech (49), other (213)
Which languages can you read fluently?	Russian (456), English (323), Ukrainian (143), Belarusian (48), German (29), other (139)	English (408), Hindi (337), Marathi (77), Gujarati (35), Sanskrit (29), other (254)	English (747), German (175), French (141), Dutch (90), Spanish (84), other (355)

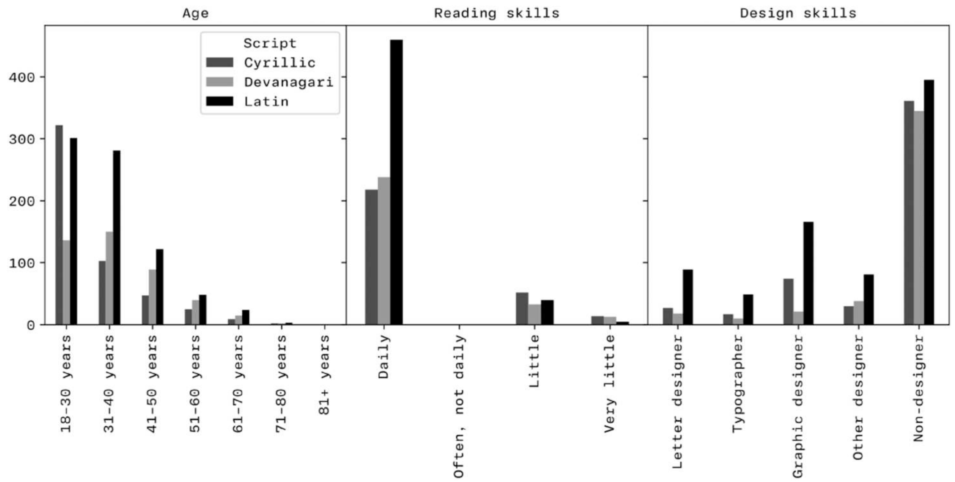


Figure 6. Charts showing participants' response counts for each script to the questions: What's your age? (Age), How often do you read books or long articles? (Reading skills), and If you are a designer, what is your discipline? (Design skills).

significant portion of non-native participants responding to the Devanagari variant. Most participants considered themselves competent readers, reading daily or often. Also, most participants are non-designers, except for the Latin script, where the call for participants started trending among designers on social media which led many of them to participate.

A total of 1721 participants took part[8] responding to 56 triplet trials each. This corresponds to a total of 96,376 trial responses across all participants. See Table 2 for the total numbers of participants per script and numbers of participants based on their nativity or fluency with respect to each script (their native/fluent languages were mapped to corresponding scripts for each language, e.g., Russian to Cyrillic or Hindi to Devanagari).

Table 2. The total numbers of participants, Native/Non-native readers, and Fluent/Non-fluent readers based on their self-reported native/fluent languages.

Script	Participants	Native	Non-native	Fluent	Non-fluent
Cyrillic	509	472	37	482	27
Devanagari	432	259	173	356	76
Latin	780	744	36	778	2

Response frequencies and inter-participant agreement

Table 3 shows the overall statistics regarding participants' responses. The relatively even spread between the theoretical minimum (.33) and maximum (1.0) of OOOO frequencies shown in Figure 7 suggests that the experimental trials covered a wide spectrum of difficulty. Furthermore, the inter-participant agreement (AC1) suggests different levels of difficulty across scripts.

Contextual pair similarity measure

The three response frequencies associated with each trial can be interpreted as measures of pair similarity within triplets. The response frequency for a particular

Table 3. Overall statistics describing participants' responses: the means with standard errors for the distribution of the OOOO responses, and AC1 calculated for all participant responses for a given script ($p < .001$ for all coefficients).

Script	OOOO mean ± SEM	AC1
Cyrillic	.69 ± .01	.56
Devanagari	.62 ± .01	.48
Latin	.72 ± .01	.60

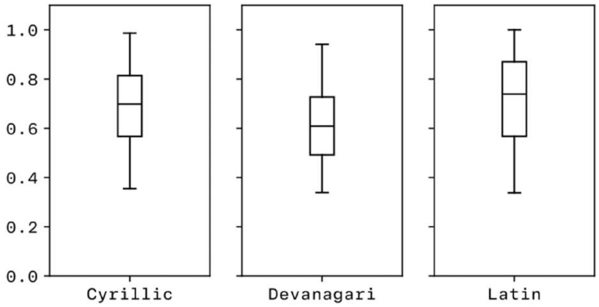
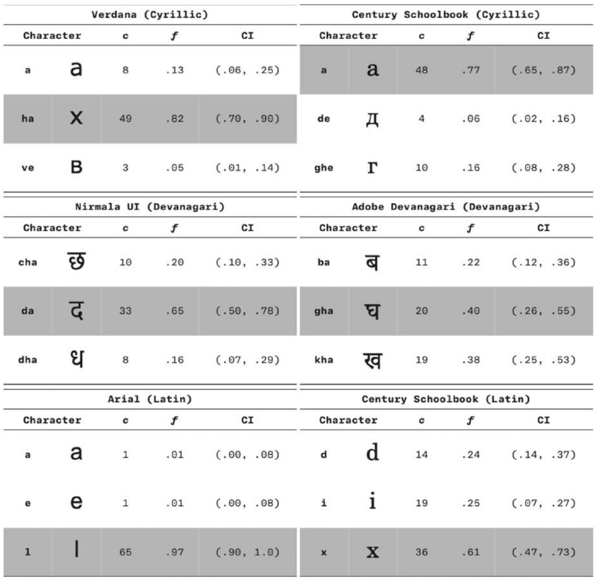


Figure 7. Box plots showing distributions of all OOOO frequencies for Cyrillic, Devanagari, and Latin.

character can be taken as a contextual similarity measure of the other two characters within the triplet set in a particular typeface. The character constitutes the context of the complementary pair of characters. This is denoted as $s(\text{a, b} \mid \text{c})_{\text{T}}$ for similarity of "a, b" in the context of "c" set in a typeface T. For example, the contextual similarity measure for the pair "d, i" in the context of "x" in the typeface Century Schoolbook (CS) is $s(\text{d, i} \mid \text{x})_{\text{CS}} = .61$ (see bottom right in Figure 8).

Results for all 1568 triplet trials (28 typeface variants with 56 trials each) that constitute the core data outcome of the experiments are available in the data repository of this project. See Figure 8 for examples. The data for the triplet trials include counts, frequencies, and confidence intervals as well as aggregated forms of the data, such as the similarity matrices described in the next section. These can be readily used by other researchers to model similarity judgements of given character triplets and pairs or to assess differences between typefaces.



Verdana (Cyrillic)

Character		c	f	CI
a	а	8	.13	(.06, .25)
ha	х	49	.82	(.70, .90)
ve	в	3	.05	(.01, .14)

Century Schoolbook (Cyrillic)

Character		c	f	CI
a	а	48	.77	(.65, .87)
de	д	4	.06	(.02, .16)
ghe	г	10	.16	(.08, .28)

Nirmala UI (Devanagari)

Character		c	f	CI
cha	छ	10	.20	(.10, .33)
da	द	33	.65	(.50, .78)
dha	ध	8	.16	(.07, .29)

Adobe Devanagari (Devanagari)

Character		c	f	CI
ba	ब	11	.22	(.12, .36)
gha	घ	20	.40	(.26, .55)
kha	ख	19	.38	(.25, .53)

Arial (Latin)

Character		c	f	CI
a	a	1	.01	(.00, .08)
e	e	1	.01	(.00, .08)
l	l	65	.97	(.90, 1.0)

Century Schoolbook (Latin)

Character		c	f	CI
d	d	14	.24	(.14, .37)
i	i	19	.25	(.07, .27)
x	x	36	.61	(.47, .73)

Figure 8. Examples of trial response data: response counts (*c*), response frequencies (*f*), and confidence intervals (CI) for trials of Cyrillic, Devanagari, and Latin with various typefaces. In the character column, character names are on the left, and small versions of the actual character images as they appeared in the experiment are on the right. The OOOO is marked with a grey background. The selection is intended to illustrate the form of the data and diversity across scripts.

Generalized character pair similarity

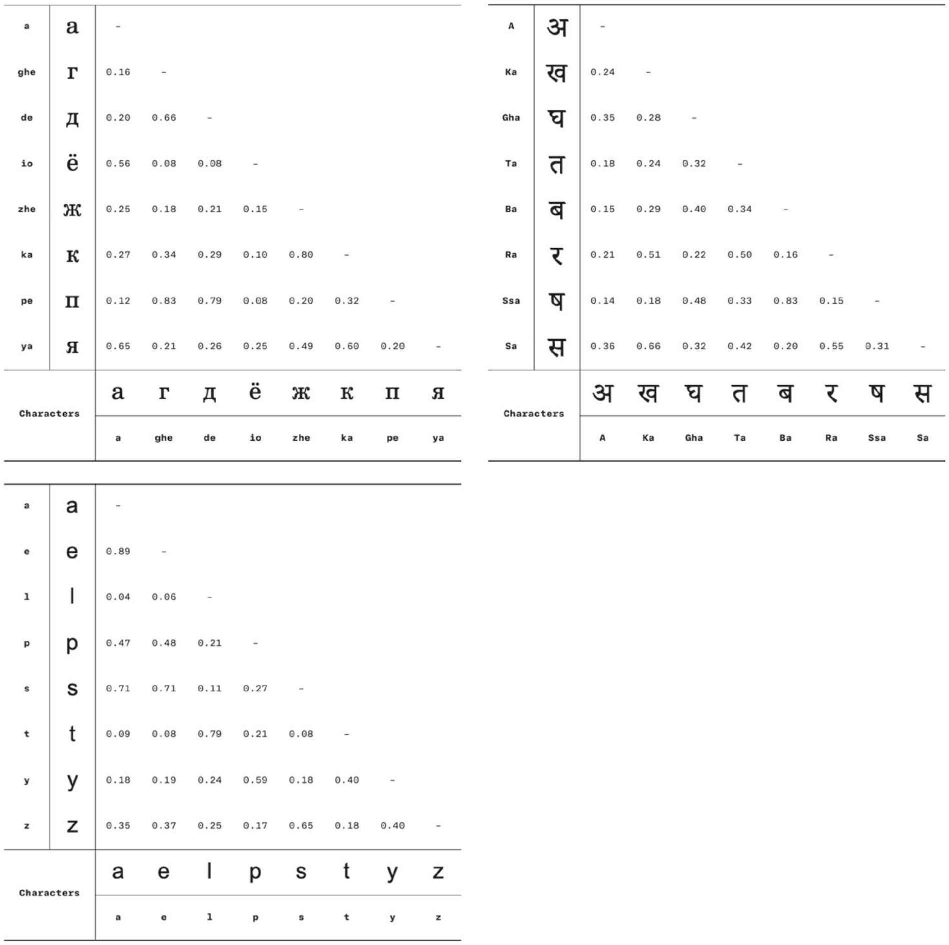
As each pair appears in multiple trials, the similarity of a particular pair can be generalized by calculating the mean of the pair's contextual similarity measures for each trial that includes the pair. In this study, the most generalized pair similarity measures are based on responses to up to six trials, i.e., six contextual characters. These generalized similarity measures for the typefaces studied can be presented in the conventional form of 8 × 8 similarity half-matrices. See three examples of similarity matrices in Figure 9 and the data repository for matrices for all 26 typefaces studied. Considering there were 61.5 (SD = 9.8) participants per typeface/trial on average, these measures are based on 61.5 × 6 = 369 responses aggregated per pair on average.

Effect of context

It is worth asking to what extent the aggregated pool of triplets, and corresponding contextual characters, influence the resulting pair-similarity measures. The box plots in Figure 10 show the distribution of variance of the similarity measures depending on the number of triplets aggregated to calculate it, indicating that there is greater variance in similarity measures based on smaller groups of triplets. This variance diminishes as the number of aggregated triplets increases, demonstrating the effect of context, which is not reflected in results from studies with individual characters or character pairs.

Comparison with other studies

The 8 × 8 matrices for specific typefaces were correlated with corresponding subsets of the similarity matrix for Arial reported by Simpson et al. (2013).[9] Table 4 shows that all matrices achieved a relatively high degree of correlation. The matrix for Arial



a	а	-							
ghe	г	0.16	-						
de	д	0.20	0.66	-					
io	ё	0.56	0.08	0.08	-				
zhe	ж	0.25	0.18	0.21	0.15	-			
ka	к	0.27	0.34	0.29	0.10	0.80	-		
pe	п	0.12	0.83	0.79	0.08	0.20	0.32	-	
ya	я	0.65	0.21	0.26	0.25	0.49	0.60	0.20	-
Characters		а	г	д	ё	ж	к	п	я
		a	ghe	de	io	zhe	ka	pe	ya

A	अ	-							
Ka	ख	0.24	-						
Gha	घ	0.35	0.28	-					
Ta	त	0.18	0.24	0.32	-				
Ba	ब	0.15	0.29	0.40	0.34	-			
Ra	र	0.21	0.51	0.22	0.50	0.16	-		
Ssa	ष	0.14	0.18	0.48	0.33	0.83	0.15	-	
Sa	स	0.36	0.66	0.32	0.42	0.20	0.55	0.31	-
Characters		अ	ख	घ	त	ब	र	ष	स
		A	Ka	Gha	Ta	Ba	Ra	Ssa	Sa

a	a	-							
e	e	0.89	-						
l	l	0.04	0.06	-					
p	p	0.47	0.48	0.21	-				
s	s	0.71	0.71	0.11	0.27	-			
t	t	0.09	0.08	0.79	0.21	0.08	-		
y	y	0.18	0.19	0.24	0.59	0.18	0.40	-	
z	z	0.35	0.37	0.25	0.17	0.65	0.18	0.40	-
Characters		a	e	l	p	s	t	y	z
		a	e	l	p	s	t	y	z

Figure 9. Examples of character pair similarity matrices based on the Cyrillic trials with the Century Schoolbook typeface (top left), Devanagari trials with the Adobe Devanagari typeface (top right), and Latin trials with the Arial typeface (bottom). The axes include the names of the selected characters as well as their rendering in the corresponding typeface. The value on the intersection for any two characters is their generalized character pair similarity measure. The diagonal entries are missing as two identical characters never occurred in the same triplet.

typeface achieved the second highest coefficient. This is to be expected as Simpson et al. based their study mainly on this typeface.

In order to assess where the data from the triplet trials offer more insights than methods based on pairs, the similarity measures from Simpson et al. (2013) were used to form count predictions corresponding to the 56 triplet trials for Arial in the present study. The chance of a character in a triplet being the OOOO is derived from the similarity measure of the other two characters from the matrix by Simpson and colleagues, denoted as *s*(a, b) for any two characters "a, b". For a triplet "a, b, c", the predicted frequencies are *f*(a, b, c) = (*s*(b, c), *s*(a, c), *s*(a, b)) and the typeface is Arial for all. These frequencies are then scaled to obtain counts with the same total per trial as the corresponding trial in the present study.

The comparison of the predicted counts with the counts from the present study has shown 79% agreement on the OOOO response (precision score: .79, recall score: .78), however the R^2 score was .36. In other words, this straightforward use of the generalized pair similarity measure explains only 36% of the variance in trial responses from the contextual similarity task in the present study.

Effects of typeface design

To explore the sensitivity of the contextual similarity task with respect to typeface design and to show the potential for finer analysis of visual features,

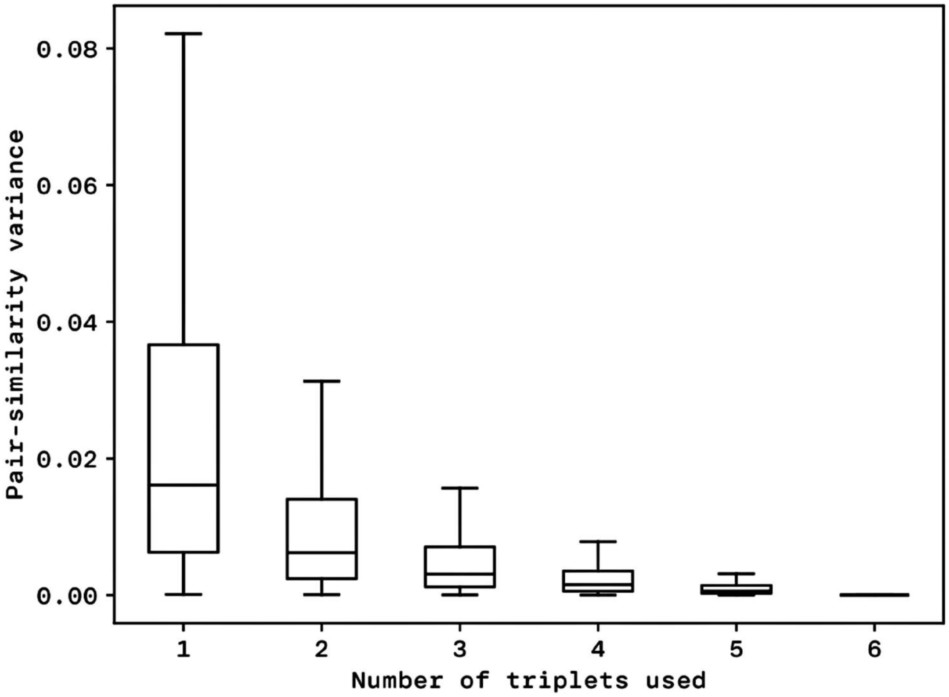


Figure 10. Each pair of characters appeared in exactly six triplets in each participant session. For each pair, these six triplets were exhaustively combined into: a set of 6 groups containing a single triplet, a set of 15 pairs of triplets, a set of 20 triplets of triplets, a set of 15 quadruplets of triplets, and a set of 6 quintuplets of triplets, and a single group of six triplets. Triplets in those groups were then aggregated to produce the generalized pair similarity measures. This was done for all character pairs studied. The box plots show the distribution of the variance in these pair similarity measures depending on the size of the groups (number of triplets) aggregated to calculate it, i.e., depending on the number of contextual characters.

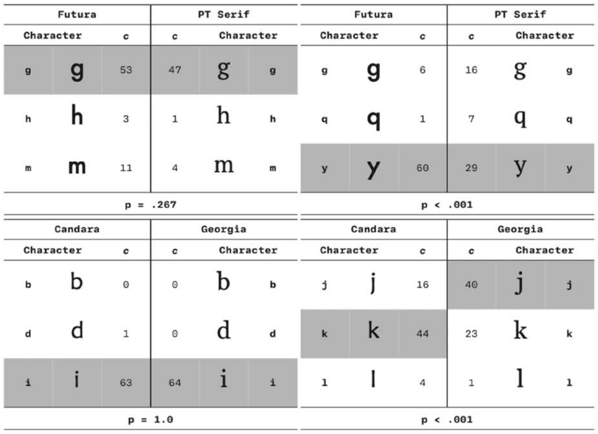
trials with the same triplet in different typefaces were compared using FET. Figure 11 shows selected examples of these comparisons. Results for all comparisons are available in the data repository.

Table 5 shows the numbers of triplets that appeared in trials with different typefaces and whether an effect of typeface design was found using FET. As testing with diverse stimuli was a priority, only four pairings of typefaces for the Latin script were tested with a large-enough triplet overlap ($m \geq 10$) to permit a meaningful comparison between different typefaces.

The results show varying degrees of differentiation between the pairs of typefaces (in relation to the

Table 4. Spearman's rank correlation coefficient of the 8×8 typeface similarity matrices with a similarity matrix reported by Simpson et al. (2013) ($p < .001$ for all coefficients, with an $\alpha = .004$ after Bonferroni correction for a family of 12 hypotheses). Only characters that occur in the present study were included in the correlation comparisons.

Similarity matrix	Spearman
Arial	.88
Calibri	.77
Cambria	.79
Candara	.62
Century Schoolbook	.81
Courier New	.81
Futura	.83
Georgia	.74
PT Sans	.89
PT Serif	.82
Times New Roman	.86
Verdana	.83



Futura			PT Serif		
Character		c	c	Character	
g	g	53	47	g	g
h	h	3	1	h	h
m	m	11	4	m	m
p = .267					

Futura			PT Serif		
Character		c	c	Character	
g	g	6	16	g	g
q	q	1	7	q	q
y	y	60	29	y	y
p < .001					

Candara			Georgia		
Character		c	c	Character	
b	b	0	0	b	b
d	d	1	0	d	d
i	i	63	64	i	i
p = 1.0					

Candara			Georgia		
Character		c	c	Character	
j	j	16	40	j	j
k	k	44	23	k	k
l	l	4	1	l	l
p < .001					

Figure 11. Juxtaposed response counts (*c*) and FET *p* for selected triplets in typefaces Futura and PT Serif (top) and Candara and Georgia (bottom). The selection is made to highlight the overall variation in the results.

Table 5. Numbers of triplets that appeared in trials with two (or more) typefaces for the Latin script. Only typeface pairings with at least ten shared triplets (m ≥ 10) are shown. The column FET shows numbers of triplets where an effect of typeface design was found using Fisher's exact test.

Typeface 1	Typeface 2	Shared triplets (*m*)	FET
Calibri	Century Schoolbook	56	5
Cambria	PT Sans	35	6
Candara	Georgia	56	17
Futura	PT Serif	56	22

selection of the triplets studied) which confirms the sensitivity of the method to typeface design effects. A few exemplary interpretations of the typeface designs and how these might have affected participants' responses are included in the Discussion.

Effects of design expertise

The following analyses explore whether there is an effect of design expertise, and whether designers are more consistent than non-designers in their similarity judgements. After preliminary analyses that did not show any regular patterns between the individual kinds of design skills as self-reported by the participants, the groups were simplified to Non-designers, Letter designers, and Designers (i.e., all designers including typographers and letter designers together).

Table 6. Numbers of participants and AC1 ($p < .001$ for all coefficients) for groups of participants divided based on their design expertise.

Script	Group	Participants (*n*)	AC1
Cyrillic	Non-designers	361	.55
	Letter designers	27	.55
	Designers	148	.57
Devanagari	Non-designers	345	.48
	Letter designers	18	.61
	Designers	87	.48
Latin	Non-designers	395	.59
	Letter designers	89	.63
	Designers	385	.60

The inter-participant agreement (AC1 in Table 6) is consistently the lowest for Non-designers, although for Cyrillic it is equal to the value reported for Letter designers and for Devanagari it is equal to the value reported for Designers. The Spearman's rank correlation and the percentage OOOO agreement (see Table 7) are consistently lower for the comparison of Non-designers to Letter designers than they are for the comparisons of Non-designers to Designers. Additionally, the MWU hypotheses tests for OOOO frequencies (see Table 7) show significant differences and marked CLES percentage between Non-designers and Letter designers in all scripts.

Effects of script fluency and nativity

Table 8 shows the participant groups divided based on their script nativity and fluency (see Table 2). Notably, the inter-participant agreement (AC1 in Table 8) is higher for Non-fluent and Non-native groups in Cyrillic and Devanagari. In contrast, the inter-participant agreement is lower for the Non-native group in Latin.

The percentage agreement and Spearman's rank correlation are lower for comparisons based on fluency (see Table 9) which suggests a greater effect of fluency than nativity for Cyrillic and Devanagari. The MWU tests for OOOO frequencies show significant results for the effect of nativity in both scripts while the effect of fluency is significant only for Devanagari. For Latin, there is not a sufficient amount of data to analyse fluency and the effect of nativity is insignificant.

Table 7. The comparisons of groups of participants based on their design expertise are described in terms of number of shared trials (m), comparison of their OOOO responses: percentage agreement and MWU for OOOO frequencies (including the corresponding p and CLES), and Spearman's rank correlation between the aggregated OOOO frequencies ($p < .001$ for all coefficients). Missing values result from an insufficient number of shared trials or participant responses for either group (m < 10, c < 5). The α = .013 after Bonferroni correction for a family of up to four comparison hypotheses per script (counted for this table and table 9). Places where the null hypothesis of equal distributions gets rejected are marked with a grey background in the p column.

			OOOO				
Script	Groups compared	Trials (*m*)	%	MWU	***p***	CLES	Spearman
Cyrillic	Non-designers vs. Letter designers	112	83.0%	3312	.006	38.3%	.70
	Non-designers vs. Designers	448	92.2%	77717	.027	45.6%	.84
Devanagari	Non-designers vs. Letter designers	56	78.6%	356	<.001	18.4%	.68
	Non-designers vs. Designers	448	81.5%	59950	.019	45.0%	.72
Latin	Non-designers vs. Letter designers	504	83.3%	68126	<.001	38.6%	.75
	Non-designers vs. Designers	672	89.1%	176415	.618	49.2%	.89

Table 8. Numbers of participants and AC1 ($p < .001$ for all coefficients) for groups of participants divided based on their script fluency and nativity.

Script	Group	Participants (*n*)	AC1
Cyrillic	Fluent	482	.55
	Non-fluent	27	.60
	Native	472	.55
	Non-native	37	.58
Devanagari	Fluent	356	.47
	Non-fluent	76	.52
	Native	259	.47
	Non-native	173	.50
Latin	Fluent	778	.60
	Non-fluent	2	–
	Native	744	.60
	Non-native	36	.57

Discussion

Demography and human decision bias

Unlike the other two scripts, there was a relatively large proportion of non-native participants responding to the Devanagari variant of the study. The reason may be that there were Indian participants who cannot be strictly considered native, although they are familiar with Devanagari from the Indian visual environment. This might have affected the distribution of the OOOO frequencies and made the trials more difficult.

Additionally, there is a clear tendency towards younger participants, possibly an effect of the study being conducted online. Consequently, care should be taken when using the data with different age groups.

According to Mueller and Weidemann (2012), successful character recognition is a result of the combined effects of perceivability, similarity, and decision bias. Human decision bias is viewed as a factor independent of the stimulus, which may include decisions based on linguistic determination of characters or personal preferences, for example. The randomization of the experimental procedures and a diverse pool of participants ensured that the effect of bias has been marginalized. In this regard, participants' diverse linguistic backgrounds are advantageous, as participants would have different biases in relation to characters' use in a language, which would reduce the chance of introducing systematic noise.

Effect of context

As shown in Figure 10, variance converges to zero as the number of aggregated triplets increases. This, perhaps expected, finding illustrates how similarity studies with character pairs oversimplify human similarity judgements. The objective of finding similarity measures for exhaustive pair combinations of the letters of a particular alphabet leads to methods that generalize with respect to context and typeface design. The goal of obtaining an exhaustive set of combinations is also problematic in terms of transferability across languages that use different sets of characters (alphabets) from a given script.

The low R2 score for the predictions modelled from the pairwise similarity matrix by Simpson et al. (2013) suggests that the data collected in the contextual similarity task are not a result of a simple transformation of some kind of general character-pair similarity values. There appear to be different decision processes involved in each kind of task.

While the contextual similarity task may seem laborious when used exhaustively, it yields deeper insights into perceptual processes. At the same time, the high correlation with the matrix by Simpson et al. (2013) supports the results from the current study and shows that the contextual similarity measure can be effectively generalized to provide compatible results.

All similarity measures presented in this paper should be considered in the context of the character selection, particular typeface, and script. The limited scope should always be kept in mind and careful consideration given when applying the measures to wider contexts.

Table 9. The comparisons of groups of participants based on their script nativity and fluency. See Table 7 for description of the statistics.

			OOOO				
Script	Groups compared	Trials (*m*)	%	MWU	***p***	CLES	Spearman
Cyrillic	Fluent vs. Non-fluent	56	87.5%	1149	.716	47.9%	.87
	Native vs. Non-native	112	92.9%	3890	<.001	36.0%	.88
Devanagari	Fluent vs. Non-fluent	448	81.5%	54804	<.001	41.1%	.71
	Native vs. Non-native	448	83.0%	60792	.004	43.9%	.77
Latin	Native vs. Non-native	168	83.3%	8867	.167	45.2%	.76

The effect of diminishing variance suggests that it may be unnecessary to aggregate exhaustive triple combinations of all characters when studying larger character sets. Instead, it may be sufficient to limit the number of triplets being aggregated and ensure diverse sampling (see section Stimuli and appendices). In Figure 10, the variance was reduced below .01 for as few as five triplets. This limit will likely depend on the number of participants, the total number of characters in a particular script, the diversity of the sample, and the script's use of similarity features.

Effects of typeface design

The diverse correlations to the matrix by Simpson et al. (2013) as shown in Table 4 readily support the effect different typeface designs may have. This can be explored in more detail thanks to the FET tests comparing results for the same triplet set in two different typefaces.

Defined by Mueller and Weidemann (2012), perceivability describes the results of constraints of the visual system and environment. In typeface design, this is usually expressed in terms of the clarity of a drawing with respect to production, display, and ergonomics, e.g., sufficient clearance between parts of characters prevents them from visually merging or complex shapes needing more overall space (Baudin, 1989; Tracy, 2003; Unger, 2018). The present study attempted to minimize the effects of perceivability by: (a) keeping the viewing conditions comfortable, and (b) focusing on conventional typeface designs intended for continuous reading. Thus, the effects of typeface design are evidenced in changing similarity relationships and participants' judgements. The judgements become more difficult when there are multiple, or an insufficient number of, criteria to choose from.

In typeface designs, similarity relationships are established through features and their configuration, i.e., presence or lack of particular features in a character. As factors of design, features may be specific shape parts such as serifs or stems or more general attributes describing openness, width, alignment, pointedness, or other notions. Hofstadter and McGraw (1998) refer to them as *parts* and *roles*, respectively. Notably, different character configurations used in typefaces lead to different perceived similarity relationships. This is most noticeable when the two configurations are largely distinct, such as in the single-–– and double-storey variants of the Latin-script letter "g" (see examples in Figure 11 where Futura uses the single-storey configuration while PT Serif uses the double-storey configuration). An effect of these variants has been formally explored by Wong et al. (2018).

Thanks to the contextual nature of the task, it is possible to hypothesize why significant differences occur. For example, for triplet "g, q, y" in the typefaces Futura and PT Serif the responses differ significantly, i.e., the typeface has an effect (top right in Figure 11). For Futura, the OOOO was clearly "y", possibly due to the fact that the construction of "y" uses diagonals while the other two characters are rounded and thus more closely related. For PT Serif, next to the "y" being the OOOO, the "g" also had a high count, possibly because it looked less related to "q" due to its more complex, double-storey variant. Note that the different variant of "g" did not have a strong effect in the triplet "g, h, m" (top left in Figure 11). This is possibly because "h, m" look very similar in both of these typefaces.

Considering the total number of recurring triplets, the cases where the typeface had an effect were not very frequent. This is likely due to the fact that all the typefaces in the studies belong to a relatively homogenous and conservatively designed group of typefaces for continuous reading. Thus, most of the characters studied share low-level feature configurations across typefaces where most differences exist in terms of treatment or salience of features, not in terms of the presence of features. For example, one might consider the differences between "b" in the Cambria typeface and "b" in the Georgia typeface subtle (see Figure 3 for examples).

The pairs of typefaces with larger numbers of shared triplets are Calibri and Century Schoolbook (56 triplets, 5 showed an effect), Cambria and PT Sans (39, 6 with effect), Candara and Georgia (56, 17 with effect), and Futura and PT Serif (56, 22 with effect). These happen to be also typefaces from two distinct genres: sans-serif and serif, respectively, in each pair. Theoretically, typefaces from the same genre would have a smaller number of triplets that show an effect. This kind of typeface comparison opens a pathway to measure perceived similarity of typefaces, perhaps on a set of key characters, to understand the visual relationships within and among different genres.

Even with a relatively homogenous group of typefaces, these results illustrate the sensitivity of the experimental method to the effects of typeface design. It is important to note that pairwise similarity measures do not provide sufficient evidence to support this level of detail in interpretation. The third, contextual character offers additional evidence to relativize the similarity within the studied pair allowing participants to make more informed decisions and rely less on memory or internal representations.

Effect of design expertise

The effect of design expertise is well-supported for Letter designers, as their OOOO response frequencies differ significantly from Non-designers. When it comes to inter-participant agreement, Designers and Letter designers are generally more consistent in their responses than Non-designers (sometimes their scores are equal), likely an effect of formal education. However, Letter designers are not always more consistent in their responses than Designers or vice versa.

Overall, the results provide support for the effect of design expertise in accordance with other studies (Dyson et al., 2016; Dyson & Stott, 2012).

Effects of script fluency and nativity

One might expect that fluent and native readers, given their experience and familiarity with a script, would have developed more consistent strategies for judging similarity. Perhaps surprisingly, the inter-participant agreement is higher among Non-fluent participants. This could be due to participants treating unfamiliar material more carefully, or additional script knowledge interfering with fluent/native participants' judgements.

In contrast, the higher inter-participant agreement for the Native group in Latin (there were no Non-fluent participants) may be due to the higher proportion of Designers taking part in the study. However, perhaps an omnipresent global script like Latin is not the right kind of stimulus when exploring an effect of fluency or nativity.

The significance of the fluency and nativity effect is confirmed for Devanagari. For Cyrillic, only the effect of nativity is significant. This may be due to a relatively small number of trials for the Fluent/Non-fluent comparison in Cyrillic. For Latin, there is not sufficient data.

There seems to be a stronger effect of fluency compared to nativity. However, considering the concepts of nativity and fluency are related and often confounded, the results point in the same direction as the findings of Gauthier et al. (2006), Wiley et al. (2016), and Wiley and Rapp (2019) that language and script expertise have an effect.

Conclusion

It is clear that data from tasks with individual characters or pairs of characters do not provide sufficient verification of theoretical models intending to be representative of perception of similarity. The context-sensitive approach is motivated by the natural attribute of visual similarity: that all similarity judgements are relative and thus depend on a context. The approach yielded more detailed data than studies with individual characters or character pairs. The contextual similarity task is cognitively simple but sufficiently challenging to elicit judgements that vary with respect to participants' background or design differences in stimuli. Eliminating the need for taxing viewing conditions helps coherence-related typeface design effects to emerge. The method was used to analyse effects of design expertise as well as script fluency and nativity, in accordance with other studies. The aggregated data from the contextual similarity task correlated with existing character pair similarity study.

Circumventing the need to test complete character sets provides new pathways to compare perceived similarity of different typefaces, which offer potential for analysis of typeface design genres and their relationship to reading performance.

The stimuli were selected from a representative pool of common typefaces across three major world scripts, and the participants provided a total of 96,376 responses. The large scope and diversity of the data supported by readers' perceptions, opens new opportunities in the exploration of internal representations, including structural, feature-based approaches and machine learning.

Views and models that consider visual similarity as contextual and typefaces as deceptively subtle, yet efficacious, can have an impact on character recognition, reading, and typeface design research by providing tools for testable critique of design decisions in typefaces, potentially connecting results of reading research to actionable design instructions.

Notes

1. For the purpose of this paper, a typeface, e.g., Arial or Times, can be defined as a notional collection of distinct styles, e.g., regular, bold, or italic. A typeface or individual styles can be implemented as a digital font.
2. This is not to say that norms do not exist, only that they describe limited portions of the whole domain.
3. To broaden out the Latin-centric perspective, the section Existing Methodologies also includes studies in other scripts. Note that the term Latin refers to the Latin script, also called the Roman alphabet. It does not refer to the Latin language.
4. The term character is used instead of letter as the latter refers solely to alphabetic units which would exclude units of other kinds of scripts, e.g., alphasyllabaries (Devanagari and other Indian scripts).
5. The data collection took place over the course of 2016.
6. For brevity, the typeface names, e.g., “Arial”, are used rather than the full style names, e.g., “Arial Regular”.
7. The typeface and the triplet might be omitted from the formula when clear from the text.
8. See appendices for exact numbers of participants for each typeface.
9. Other authors either reported on uppercase letters only, did not study scripts used in this study, or worked with very small numbers of participants ($n \leq 30$).

Acknowledgements

This research was part of the author’s work towards the degree of Doctor of Philosophy at the Department of Typography & Graphic Communication, University of Reading (UK). It was supervised by Dr Mary C. Dyson, Dr Fiona G. E. Ross, and Keith Chi-Hang Tam.

Data and code

The complete data from all trial responses in their raw and aggregated form are made available in an Open Science Framework repository at https://osf.io/de4n7/ (DOI: 10.17605/OSF.IO/DE4N7). Also available are the study website and Jupyter notebooks documenting the statistical analyses and preprocessing.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research was funded by the UK Arts and Humanities Research Council under the auspices of the Design Star Doctoral Training Centre.

Consent to participate and consent for publication

Informed consent was obtained from all individual participants included in the study.

Statistics

The statistics in this paper were calculated using Python programming language with SciPy, Pandas, statmodels, Pingouin, irrCAC, matplot, and RPy2 libraries. The analysis plan was not pre-registered.

Ethics approval

The study was reviewed by the Research Ethics Committee of the University of Reading and given a favourable ethical opinion for conduct in February 2016.

ORCID

David Březina http://orcid.org/0009-0007-4410-0133

References

Alexeeva, S. (2024). Parafoveal letter identification in Russian: Confusion matrices based on error rates. *Behavior Research Methods*, 8567–8587. https://doi.org/10.3758/s13428-024-02492-3

Baudin, F. (1989). *How typography works: (And why it is important)*. Lund Humphries Publishers.

Beier, S., & Larson, K. (2013). How does typeface familiarity affect reading performance and reader preference. *Information Design Journal*, *20*(1), 16–31. https://doi.org/10.1075/idj.20.1.02bei

Biederman, I. (1987). Recognition-by-Components: A theory of human image understanding. *Psychological Review*, *94*(2), 115–147. https://doi.org/10.1037/0033-295X.94.2.115

Blasi, D. E., Henrich, J., Adamou, E., Kemmerer, D., & Majid, A. (2022). Over-reliance on English hinders cognitive science. *Trends in Cognitive Sciences*, *26*(12), 1153–1170. https://doi.org/10.1016/j.tics.2022.09.015

Boles, D. B., & Clifford, J. E. (1989). An upper- and lowercase alphabetic similarity matrix, with derived generation similarity values. *Behavior Research Methods, Instruments & Computers*, *21*(6), 579–586. https://doi.org/10.3758/BF03210580

Boudelaa, S., Perea, M., & Carreiras, M. (2020). Matrices of the frequency and similarity of Arabic letters and allographs. *Behavior Research Methods*, *52*(5), 1893–1905. https://doi.org/10.3758/s13428-020-01353-z

Bouma, H. (1971). Visual recognition of isolated lower case letters. *Vision Research*, *11*(5), 459–474. https://doi.org/10.1016/0042-6989(71)90087-3

Březina, D. (2021). Character complexity and redundancy in writing systems over human history [review]. *Design Regression*. https://designregression.com/review/character-complexity-and-redundancy-in-writing-systems-over-human-history

Chang, L. Y., Chen, Y. C., & Perfetti, C. A. (2018). Graphcom: A multidimensional measure of graphic complexity applied to 131 written languages. *Behavior Research Methods*, *50*(1), 427–449. https://doi.org/10.3758/s13428-017-0881-y

Changizi, M. A., & Shimojo, S. (2005). Character complexity and redundancy in writing systems over human history. *Proceedings of the Royal Society B: Biological Sciences*, *272*(1560), 267–275. https://doi.org/10.1098/rspb.2004.2942

Changizi, M. A., Zhang, Q., Ye, H., & Shimojo, S. (2006). The structures of letters and symbols throughout human history are selected to match those found in objects in natural scenes. *The American Naturalist*, *167*(5), E117–E139. https://doi.org/10.1086/502806

Clopper, C. J., & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, *26*(4), 404–413. https://doi.org/10.1093/biomet/26.4.404

Courrieu, P., Farioli, F., & Grainger, J. (2004). Inverse discrimination time as a perceptual distance for alphabetic characters. *Visual Cognition*, *11*(7), 901–919. https://doi.org/10.1080/13506280444000049

Dehaene, S. (2010). *Reading in the brain*. Penguin.

Dyson, M. C. (2019). *Legibility: How and why typography affects ease of reading*. Centro de Estudios Avanzados de Diseño.

Dyson, M. C., & Stott, C. (2012). Characterizing typographic expertise: Do we process typefaces like faces. *Visual Cognition*, *20*(9), 1082–1094. https://doi.org/10.1080/13506285.2012.722568

Dyson, M. C., Tam, K., Leake, C., & Kwok, B. (2016). How does expertise contribute to the recognition of Latin and Chinese characters?. In M. C. Dyson, & C. Y. Suen (Eds.), *Digital fonts and reading* (pp. 193–208). World Scientific Publishing.

Fiset, D., Blais, C., Arguin, M., Tadros, K., Éthier-Majcher, C., Bub, D., & Gosselin, F. (2009). The spatio-temporal dynamics of visual letter recognition. *Cognitive Neuropsychology*, *26*(1), 23–35. https://doi.org/10.1080/02643290802421160

Fiset, D., Blais, C., Éthier-Majcher, C., Arguin, M., Bub, D., & Gosselin, F. (2008). Features for identification of uppercase and lowercase letters. *Psychological Science*, *19*(11), 1161–1168. https://doi.org/10.1111/j.1467-9280.2008.02218.x

Fisher, R. A. (1922). On the interpretation of χ2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society*, *85*(1), 87–94. https://doi.org/10.2307/2340521

Garces, E., Agarwala, A., Gutierrez, D., & Hertzmann, A. (2014). A similarity measure for illustration style. *ACM Transactions on Graphics (TOG)*, *33*(4), 93. https://doi.org/10.1145/2601097.2601131

Gauthier, I., Wong, A. C., Hayward, W. G., & Cheung, O. S. (2006). Font tuning associated with expertise in letter perception. *Perception*, *35*(4), 541–559. https://doi.org/10.1068/p5313

Geyer, L. H. (1977). Recognition and confusion of the lowercase alphabet. *Perception & Psychophysics*, *22*(5), 487–490. https://doi.org/10.3758/BF03199550

Gibson, E. P., Gibson, J. J., Pick, A. D., & Osser, H. (1962). A developmental study of the discrimination of letter-like forms. *Journal of Comparative and Physiological Psychology*, *55*(6), 897–906. https://doi.org/10.1037/h0043190

Gill, E. (1988). *An essay on typography*. D. R. Godine. (Original work published 1931).

Gilmore, G. C., Hersh, H., Caramazza, A., & Griffin, J. (1979). Multidimensional letter similarity derived from recognition errors. *Perception & Psychophysics*, *25*(5), 425–431. https://doi.org/10.3758/BF03199852

Grainger, J., Dufau, S., & Ziegler, J. C. (2016). A vision of reading. *Trends in Cognitive Sciences*, *20*(3), 171–179. https://doi.org/10.1016/j.tics.2015.12.008

Grainger, J., Rey, A., & Dufau, S. (2008). Letter perception: From pixels to pandemonium. *Trends in Cognitive Sciences*, *12*, 381–387. https://doi.org/10.1016/j.tics.2008.06.006

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, *61*, 29–48. https://doi.org/10.1348/000711006X126600

Higuchi, H., & Kobayashi, T. (2022). Letter visual similarity of Japanese hiragana and katakana based on reaction times. *Current Psychology*, 1–10. https://doi.org/10.1007/s12144-021-02664-w

Hinton, G. E., Williams, C. K. I., & Revow, M. D. (1992). Adaptive Elastic Models for Hand-Printed Character Recognition. In *Advances in Neural Information Processing Systems 4 (NIPS 1991)* (pp. 512–519). Morgan-Kaufmann. https://papers.nips.cc/paper_files/paper/1991/hash/df877f3865752637daa540ea9cbc474f-Abstract.html

Hofstadter, D. R., & McGraw, G. (1998). Letter spirit: Esthetic perception and creative play in the rich microcosm of the Roman alphabet. In *Fluid concepts & creative analogies: Computer models of the fundamental mechanisms of thought* (pp. 407–466). Penguin Books.

Hollis, G., & Westbury, C. (2018). When is best-worst best? A comparison of best-worst scaling, numeric estimation, and rating scales for collection of semantic norms. *Behavior Research Methods*, *50*(1), 115–133. https://doi.org/10.3758/s13428-017-1009-0

Kaiho, H. (1970). Similarity dimensions and legibility of Kata-Kana letters. *The Japanese Journal of Psychology*, *40*(6), 337–340. https://doi.org/10.4992/jjpsy.40.337

Kelly, G. A. (1955). *The psychology of personal constructs*. W. W. Norton.

Kuennapas, T., & Janson, A. J. (1969). Multidimensional similarity of letters. *Perceptual and Motor Skills*, *28*(1), 3–12. https://doi.org/10.2466/pms.1969.28.1.3

Lally, C., & Rastle, K. (2023). Orthographic and feature-level contributions to letter identification. *Quarterly Journal of Experimental Psychology*, *76*(5), 1111–1119. https://doi.org/10.1177/17470218221106155

Lanthier, S. N., Risko, E. F., Stolz, J. A., & Besner, D. (2009). Not all visual features are created equal: Early processing in letter and word recognition. *Psychonomic Bulletin & Review*, *16*(1), 67–73. https://doi.org/10.3758/pbr.16.1.67

Liu, L., & Arditi, A. (2001). How crowding affects letter confusion. *Optometry & Vision Science*, *78*(1), 50–55. https://doi.org/10.1097/00006324-200101010-00014

Lun, Z., Kalogerakis, E., & Sheffer, A. (2015). Elements of style: Learning perceptual shape style similarity. *ACM Transactions on Graphics*, *34*(4), 1–14. https://doi.org/10.1145/2766929

Marcet, A., & Perea, M. (2017). Is nevtral NEUTRAL? Visual similarity effects in the early phases of written-word recognition. *Psychonomic Bulletin & Review*, *24*(4), 1180–1185. https://doi.org/10.3758/s13423-016-1180-9

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review*, *88*(5), 375. https://doi.org/10.1037/0033-295X.88.5.375

McFee, B., & Lanckriet, G. (2011). Learning multi-modal similarity. *Journal of Machine Learning Research*, *12*(15), 491–523. https://doi.org/10.1145/1273496

Meletis, D. (2020). *The Nature of Writing*. https://doi.org/10.36824/2020-meletis

Mueller, S. T., & Weidemann, C. T. (2012). Alphabetic letter identification: Effects of perceivability, similarity, and bias. *Acta Psychologica*, *139*(1), 19–37. https://doi.org/10.1016/j.actpsy.2011.09.014

Palmer, S. E. (1999). *Vision science : Photons to phenomenology*. MIT Press.

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/10.1016/j.visres.2006.04.023

Pelli, D. G., Farell, B., & Moore, D. C. (2003). The remarkable inefficiency of word recognition. *Nature*, *423*(6941), 752–756. https://doi.org/10.1038/nature01516

Pelli, D. G., Tillman, K. A., Freeman, J., Su, M., Berger, T. D., & Majaj, N. J. (2007). Crowding and eccentricity determine reading rate. *Journal of Vision*, *7*(2), 1–36. https://doi.org/10.1167/7.2.20

Perea, M., Baciero, A., & Marcet, A. (2021). Does a mark make a difference? Visual similarity effects with accented vowels. *Psychological Research*, *85*(6), 2279–2290. https://doi.org/10.1007/s00426-020-01405-1

Petit, J.-P., & Grainger, J. (2002). Masked partial priming of letter perception. *Visual Cognition*, *9*(3), 337–353. https://doi.org/10.1080/13506280042000207

Phillips, J. R., Johnson, K. O., & Browne, H. M. (1983). A comparison of visual and two modes of tactual letter resolution. *Perception & Psychophysics*, *34*(3), 243–249. https://doi.org/10.3758/BF03202952

Pick, A. D. (1965). Improvement of visual and tactual form discrimination. *Journal of Experimental Psychology*, *69*(4), 331–339. https://doi.org/10.1037/h0021772

Podgorny, P., & Garner, W. R. (1979). Reaction time as a measure of inter- and intraobject visual similarity: Letters of the alphabet. *Perception & Psychophysics*, *26*(1), 37–52. https://doi.org/10.3758/bf03199860

Reich, L. N., & Bedell, H. E. (2000). Relative legibility and confusions of letter acuity targets in the peripheral and central retina. *Optometry and Vision Science*, *77*(5), 270–275. https://doi.org/10.1097/00006324-200005000-00014

Rosa, E., Perea, M., & Enneson, P. (2016). The role of letter features in visual-word recognition : Evidence from a delayed segment technique. *Acta Psychologica*, *169*, 133–142. https://doi.org/10.1016/j.actpsy.2016.05.016

Rothlein, D., & Rapp, B. (2017). The role of allograph representations in font-invariant letter identification. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(7), 1411. https://doi.org/10.1037/xhp0000384

Sanocki, T. (1987). Visual knowledge underlying letter perception: Font-specific, schematic tuning. *Journal of Experimental Psychology: Human Perception and Performance*, *13*(2), 267–278. https://doi.org/10.1037//0096-1523.13.2.267

Sanocki, T., & Dyson, M. C. (2012). Letter processing and font information during reading: Beyond distinctiveness, where vision meets design. *Attention, Perception, & Psychophysics*, *74*(1), 132–145. https://doi.org/10.3758/s13414-011-0220-9

Schultz, M., & Joachims, T. (2003). Learning a distance metric from relative comparisons. In S Thun, L Saul, & B Schölkopf (Eds.), *Advances in neural information processing systems 16 (NIPS 2003)* (pp. 41–48). https://proceedings.neurips.cc/paper/2003/hash/d3b1fb02964aa64e257f9f26a31f72cf-Abstract.html.

Simpson, I. C., Mousikou, P., Montoya, J. M., & Defior, S. (2013). A letter visual-similarity matrix for Latin-based alphabets. *Behavior Research Methods*, *45*(2), 431–439. https://doi.org/10.3758/s13428-012-0271-4

Smeijers, F. (1996). *Counterpunch : Making type in the sixteenth century, designing typefaces now* (2nd ed.). Hyphen Press.

Southall, R. (1986). Shape and appearance in typeface design. *Protext*, *III*, 75–86.

Southall, R. (2005). *Printer's type in the twentieth century : Manufacturing and design methods*. Oak Knoll Press.

The Unicode Consortium. (2023). *The Unicode Standard (Version 15.1.0)*. https://www.unicode.org/versions/Unicode15.1.0/

Townsend, J. T. (1971). Theoretical analysis of an alphabetic confusion matrix. *Perception & Psychophysics*, *9*(1), 40–50. https://doi.org/10.3758/bf03213026

Tracy, W. (2003). *Letters of credit : A view of type design* (First softcover ed.). D. R. Godine. https://doi.org/10.1086/pbsa.83.2.24303739

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327–352. https://doi.org/10.1037/0033-295X.84.4.327

Unger, G. (2018). *Theory of type design*. Nai010 Publishers.

Walker, P. (2008). Font tuning: A review and new experimental evidence. *Visual Cognition*, *16*(8), 1022–1058. https://doi.org/10.1080/13506280701535924

Wang, X., Kitani, K. M., & Hebert, M. (2016). Contextual Visual Similarity. *arXiv*. https://doi.org/10.48550/arXiv.1612.02534

Whistler, K., & Freytag, A. (2022). *Unicode® Technical Report #17: Unicode character encoding model*. http://unicode.org/reports/tr17/#CharactersVsGlyphs

White, S. J., Johnson, R. L., Liversedge, S. P., & Rayner, K. (2008). Eye movements when reading transposed text: The importance of word-beginning letters. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(5), 1261–1276. https://doi.org/10.1037/0096-1523.34.5.1261

Wiley, R. W., & Rapp, B. (2019). From complexity to distinctiveness: The effect of expertise on letter perception. *Psychonomic Bulletin & Review*, *26*(3), 974–984. https://doi.org/10.3758/s13423-018-1550-6

Wiley, R. W., Wilson, C., & Rapp, B. (2016). The effects of alphabet and expertise on letter perception. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(8), 1186–1203. https://doi.org/10.1037/xhp0000213

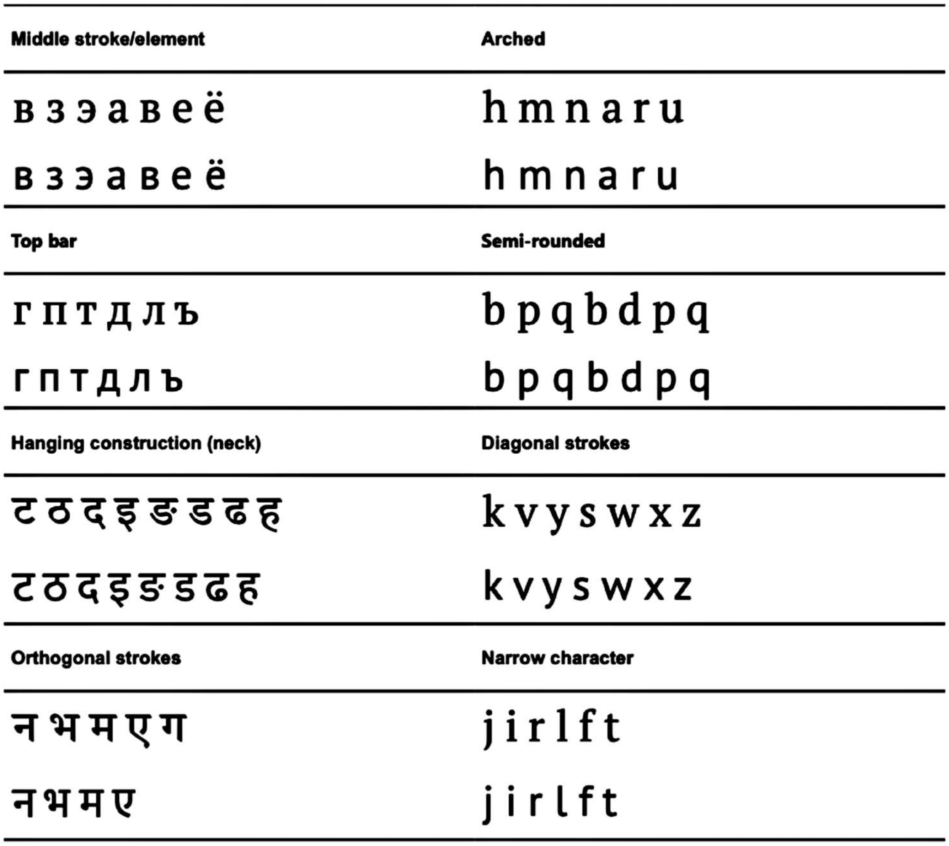
Wong, K., Wadee, F., Ellenblum, G., & McCloskey, M. (2018). The devil's in the g-tails: Deficient letter-shape knowledge and awareness despite massive visual experience. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(9), 1324–1335. https://doi.org/10.1037/xhp0000532

Appendices

Selection of characters for each typeface variant of the experiments

Eight characters for each typeface-script combination were selected systematically (see figure A1). The goal was to represent characters as evenly as possible across all experiment variants and to combine highly similar characters as well as highly dissimilar characters in each selection. The selection process was based on subjective, observed visual attributes.

The Latin script was represented with selections from the lowercase letters of the English alphabet. The Cyrillic script was represented with selections from the lowercase letters of the Russian alphabet (including a single character with a diacritical mark). Lowercase letters were preferred as they appear more frequently and, compared to uppercase letters, they have a greater effect on the visual appearance of long-form texts. The Devanagari script was represented with selections of the consonantal characters from the basic Hindi syllabary. Although essential for the correct use of the script, vowel characters, vowel marks and mark combinations, and consonantal clusters (conjuncts) were not studied as these occur less frequently in contemporary Hindi and would increase the scope of potential combinations.



Middle stroke/element	Arched
в з э а в е ё	h m n a r u
в з э а в е ё	h m n a r u
Top bar	**Semi-rounded**
г п т д л ъ	b p q b d p q
г п т д л ъ	b p q b d p q
Hanging construction (neck)	**Diagonal strokes**
ट ठ द इ ङ ड ढ ह	k v y s w x z
ट ठ द इ ङ ड ढ ह	k v y s w x z
Orthogonal strokes	**Narrow character**
न भ म ए ग	j i r l f t
न भ म ए	j i r l f t

Figure A1. Examples of visual attributes used for selection of characters: each attribute is presented with a group of characters in a high-contrast (first row) and low-contrast typeface (second row). A single character could have multiple attributes, i.e., it could occur in multiple attribute groups.

Systematic grouping of characters into the attribute groups could potentially introduce bias by pre-selecting combinations of characters that would be easy to judge according to the observed attributes. This risk was reduced by using a different selection of characters for most of the typefaces and combining the characters exhaustively to produce a variety of combinations.

Pilot

The experiment was piloted using 23 triplets of characters from a single typeface. The website was a prototype, somewhat different in design from the website finally used. There were six participants who were informally interviewed afterwards. It took them 2–3 minutes to work through all triplets. The pilot showed that participants tend to agree on the more obvious triplet trials. They did not have a problem understanding the task and using the website. Some participants with design knowledge focused on fine details. Some thought the task was to spot a letter from a different typeface within each triplet. This ultimately led to rephrasing of the instructions to discourage participants from focusing on details.

Details regarding participants in the main study

The website traffic analytic tools reported that a total of 3428 people visited the websites and 1475 successfully completed the experiment and submitted their responses. The approximate dropout rate was 57%. In total, 49.4% of visitors used a desktop computer, 41.6% used a mobile phone, 8.5% used a tablet, with the remainder of 0.5% using an unknown device. Due to the nature of traffic analytics, these numbers should be considered estimates.

The exact numbers of participants that completed the experiment are reported in table A1.

Table A1. The numbers of participants for each typeface variant of the experiment.

	Script / Participants		
Typeface	Cyrillic	Devanagari	Latin
Arial	69		67
Century Schoolbook	62		59
Courier New	67		85
Georgia	75		64
PT Sans	53		50
PT Serif	50		52
Times New Roman	73		61
Verdana	60		63
Calibri			77
Cambria			74
Candara			61
Futura			67
Adobe Devanagari		50	
Devanagari MT		55	
Ek Mukta		58	
ITF Devanagari		48	
Kohinoor Devanagari		45	
Lohit Devanagari		63	
Murty Hindi		62	
Nirmala UI		51	

Distinctiveness tables

Beyond the pairwise similarity matrices, the data can be also aggregated character-wise, each response frequency for a particular character averaged across all triplets for a single typeface. The result measures how often a character is considered the odd one out and can be loosely interpreted as a measure of the character's *distinctiveness* from the other seven characters in a variant of the experiment with a particular typeface. The distinctiveness tables are available in the data repository for potential comparison with single-character recognition studies by other authors.