ISSN 0022-2224

Published continuously
since 1967.

Visible Language

56 . 3

the journal of visual communication research

december. 2022

Visible Language

the journal of visual communication research

Thiessen et al.

08 – 31

how the complexity of letterforms affects the reading process, with special attention to the use of a specially designed typeface to control the number of variables and improve research validity

Tjung et al.

32 – 57

a study to understand the features of learning apps required for mobile gamification

Kudo

58 – 85

a study of the most beneficial changes to pictograms to facilitate correct comprehension by people with intellectual disabilities, including recommendations for both design and research methods

red

red

red

red

Before there was reading there was seeing.

People navigate the world and probe life's meaning through visible language. *Visible Language* has been concerned with ideas that help define the unique role and properties of visual communication. A basic premise of the journal has been that visual design is a means of communication that must be defined and explored on its own terms. This journal is devoted to enhancing people's experience through the advancement of research and practice of visual communication.

Published three time a year, in April, August, and December

website:

http://visiblelanguagejournal.com

send editorial correspondence to:

Mike Zender, *Editor*
College of Design, Architecture, Art, and Planning
University of Cincinnati
PO Box 210016
Cincinnati, OH 45221-0016
mike.zender@uc.edu

Professor Matthew Wizinsky, *Associate Editor*
matthew.wizinsky@uc.edu

Professor Muhammad Rahman, *Assistant editor*
rahmanmd@ucmail.uc.edu

Sharon Poggenpohl, *Editor Emeritus*

Merald Wrolstad, *Founder*

direct all subscription inquiries to:
Carly Truitt - pubsvc.tsp@sheridan.com

University of Cincinnati, School of Design, UCPress - *Publisher*
© Copyright 2022 by the University of Cincinnati

Subscription Rates (effective 02/15/21)

Individual	
1 year (print only)	$55.00
1 year (e-only)	$49.00
2 year (print only)	$110.00
2 year (e-only)	$95.00
Institutional	
1 year (e-only)	$196.00
1 year (print only)	$220.00
1 year (print and e)	$250.00

Please direct all subscription inquiries to:
Carly Truitt - pubsvc.tsp@sheridan.com

Prepayment is required. Make checks payable to University of Cincinnati *Visible Language* in U.S. currency only, foreign banks need a U.S. correspondent bank.

ISSN 0022-2224
Published continuously since 1967.

Back Copies

A limited number of nearly all back numbers is available. The journal website at http://visiblelanguagejournal.com is searchable and lists all issues, past article PDFs, contents, and abstracts.

Copyright Information

Authorization to photocopy items for internal or personal use, or for libraries and other users registered with the Copyright Clearance Center (CCC) Transactional Reporting Service, provided that the base fee of $1.00 per article, plus .10 per page is paid directly to:

CCC
21 Congress Street
Salem, Massachusetts 01970
Telephone 508.744.3350
0022-22244/86 $1.00 plus .10

UNIVERSITY OF
Cincinnati

Visible
Language

56 — 3

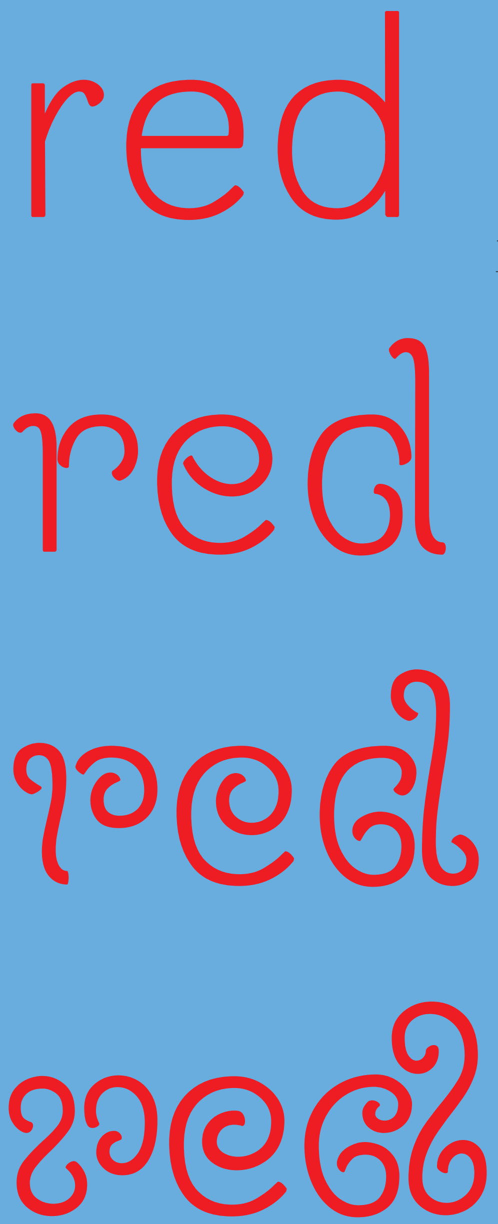
_2022

_december

Advisory Board

Naomi Baron – *The American University, Washington, D.C.*
Michael Bierut – *Pentagram, New York, NY*
Ann Bessemans – *Hasselt University & PXL-MAD (Media, Arts, Design, School of Arts, Hasselt, Belgium*
Charles Bigelow – *Type designer*
Matthew Carter – *Carter & Cone Type, Cambridge, MA*
Keith Crutcher – *Cincinnati, OH*
Meredith Davis – *Emerita Alumni Distinguished Graduate Professor, Department of Graphic and Industrial Design, College of Design / NC State University*
Mary Dyson – University of Reading, UK
Jorge Frascara – *University of Alberta, Canada*
Ken Friedman – *Chair Professor of Design Innovation Studies, College of Design and Innovation, Tongji University*
Michael Golec – *School of the Art Institute of Chicago, Chicago, IL*
Judith Gregory – *University of California-Irvine, Irvine, CA*
Dori Griffin – *University of Florida School of Art + Art History*
Kevin Larson – *Microsoft Advanced Reading Technologies*
Aaron Marcus – *Aaron Marcus & Associates, Berkeley, CA*
Tom Ockerse – *Rhode Island School of Design, Providence, RI*
Sharon Poggenpohl – *Estes Park, CO*
Michael Renner – *The Basel School of Design, Visual Communication Institute, Academy of Art and Design, HGK FHNW*
Stan Ruecker – *IIT, Chicago, IL*
Katie Salen – *DePaul University, Chicago, IL*
Peter Storkerson – *Champaign, IL*
Karl van der Waarde – *Swinburne University of Technology, Melbourne*
Mike Zender – *University of Cincinnati, Cincinnati, OH*

Contents december 2022



Effect of Typeface Complexity on Automatic Whole-Word Reading Processes

Myra Thiessen

Hannah Keage

Indae Hwang

Jack Astley

Sofie Beier

Abstract

Visually complex typefaces require more cognitive effort to process, which can impact reading efficiency, and have been associated with disfluency effects. Since our environments may include an increasing range of demanding reading scenarios—to which we are expected to respond, sometimes with speed and accuracy—it is important to develop an understanding of how reading proficiency may be affected as a result. With a focus on how automatic reading processes may be affected, this study explores the impact of typeface complexity, determined by stroke length and systematically measured using perimetric complexity, by using the well-known Stroop Color and Word Test. We show that automatic whole-word reading can be negatively affected by typefaces with extremely complex features, but that moderately complex typefaces have little effect. This suggests that hard-to-read typefaces do impair word reading (i.e., they are disfluent) but that skilled readers are able to tolerate a high degree of complexity. It also highlights the utility of cognitive tests for identifying typefaces that are difficult to read.

Keywords:

typeface complexity;
font design;
Stroop Test;
automatic reading;
disfluency effect

Introduction

Varying the visual complexity of letterforms is associated with a novelty effect like that seen in font tuning (Gauthier et al., 2006; Sanocki, 1987, 1988; Sanocki & Dyson, 2012), which states that readers must adjust and learn new shapes of typefaces that are unfamiliar or novel in style. It can be argued that the further novel letter shapes stray from the neutral letter skeleton, the more learning a reader must do, which may impact reading efficiency, resulting in a "bottleneck" in visual perception (Bernard & Chung, 2011; Pelli et al., 2006). Since there are indications that embellishments like exaggerated swash[1] styles inhibit letter identification (Beier et al., 2017), this study investigates the automatic processing of words and the effect of added visual complexity defined by increasing the stroke length of letterforms, inspired by swash embellishments. Automatic reading processes refer to reading actions that are perceived to be effortless by skilled readers and include activities that have been developed over time through continued practice. This includes actions such as letter recognition and word reading and tends to demand only minimal attention and cognitive load[2] (Walczyk, 2000). In this study, we examine automatic whole-word reading processes by drawing on a standard Stroop Test paradigm.

Word Identification and Font Disfluency

The effect that swash embellishments and exaggerated letter strokes have on reading can be informed by existing letter and word identification literature. There is a general consensus in cognitive neuroscience that feature detection describes the primary means for letter identification, purporting that readers access specific unique and identifiable parts of letters in a hierarchical manner, rather than drawing on the whole letter, during reading (Grainger et al., 2008). Letters and words are further recognized through parallel hierarchical processes distinguishing letter features, whole letters, and words (Coltheart et al., 2001; Reichle, 2020). Further, it is important to note that a number of experiments attempt to understand letter recognition by identifying the essential features readers rely upon by using techniques such as degrading or removing parts of the stimuli letters. However, several discrepancies can be observed with the results. These experiments have shown that eliminating the middle portion of letters resulted in the worst performances compared to eliminating the junctions and stroke terminations (Petit & Grainger, 2002; Rosa et al., 2016); conversely, others state that it is worse to eliminate the junctions than the midsections (Lanthier et al., 2009). Others again have found that removing stroke terminations created the most difficulty for letter recognition (Fiset et al., 2008).

The discrepancy across these experiments is concerning, but not uncommon. It may be due to individual differences across readers, which is evidenced in Dyson and Brezina (2021) who showed that individuals with typographic expertise are more sensitive to typographic variation than those who are untrained, and that this can affect their judgements of learning. It may also be due to stylistic differences in the typefaces used to develop test materials. The visual and stylistic properties of typefaces may play a role in the outcomes reported in legibility studies, particularly when typefaces originating from broad typeface categories are compared, e.g., serif and sans serif or monospaced and non-monospaced. Thus, isolating and accounting for stylistic typographic variables are important considerations in legibility experiments since tests that draw on a range of different typeface families may introduce variables that are unaccounted for in the results.

It is suggested that the ease with which letters and words can be recognized is affected by the clarity and visual simplicity of the font design. By overlaying the same letters in a range of common fonts, Figure 1 shows how a "neutral" or "standard" letter skeleton can be identified. Letters that closely align to the neutral skeleton are easier to recognized, or are more fluent, because their shape draws on familiar, idealized, or essential letter shapes (Beier et al., 2017; Frutiger, 1989). Conversely, font styles that include deviations from the neutral skeleton, such as those with added embellishments, have been shown to inhibit letter recognition (Beier et al., 2017; Pelli et al., 2006) as a result of their visual complexity. The visual complexity of a letter can be determined by measuring perimetric complexity, which is the measurement of the perimeter of a character (inside and outside). The perimeter total is squared and then divided by the "fill" or "ink" area (Pelli et al., 2006). Letters that have exaggerated stroke lengths, like those with swash embellishments, are likely to have a higher perimetric measurement and are, therefore, considered to be more complex.

---

[1] A swash is an embellishment, flourish, or decorative element, sometimes seen on fonts like scripts. Swash embellishments are added at letter terminals and may include an exaggerated serif or tail. The tested letterforms in this study do not contain swash embellishments in the traditional sense since we have increased the stroke length throughout the letterform; however, we are inspired by the added visual complexity of these decorative elements and interested in their impact on legibility.

[2] Counterpart to automatic reading processes are controlled reading processes. These are more cognitively demanding and include more complex activities, like decoding an unfamiliar word and integrating meaning, and also require conscious attention (Walczyk, 2000).

Figure 1

Series of overlapping typefaces demonstrating the neutral, common, or familiar letter skeleton. Fonts used are both serif and sans serif and are a representation of commonly used varieties. They are: Baskerville, Helvetica, Minion Pro, Myriad Pro, Times New Roman, and Verdana.

The visual complexity of a typeface is likely one factor that contributes to disfluency effects in reading, which is described as the perceived effort needed to complete a reading task (Oppenheimer, 2008). Studies show complex typefaces attract more cognitive effort on both perceptual and higher-order levels (Keage et al., 2014; Thiessen et al., 2015), but whether this extra effort is desirable for reading related tasks continues to be debated (Diemand-Yauman et al., 2011; Geller et al., 2020; Taylor et al., 2020; Thiessen et al., 2020). It is easy to see the importance of this discussion in the context of functional reading and the impact that environmental distractions have on attention. It is argued that a better understanding of which typographic features disrupt automatic reading may improve outcomes for these more cognitively demanding tasks. Complex reading scenarios are becoming more commonplace and readers are expected to interact with displays that "allow information to be presented to a driver without necessitating glances away from the roadway, a security camera might provide location information over feed, or a display might deliver notification information superimposed over a user-selected background" (Sawyer et al., 2020, p. 865).

The prospect of receiving information without looking away from the road while driving has the appeal of efficiency and safety; however, this may be very far from the truth and a reader's capacity to process such visually complex information could be severely compromised. Sawyer and colleagues (2020) showed that the level of complexity of both background information and the typeface layered over top can impact legibility in glanceable reading scenarios (e.g., driving). In fact, techniques that typographers may rely on to improve legibility when layering type over an image, like adding an outline, was shown to reduce legibility compared to less visually complex techniques, like adding a drop shadow. Since more visually complex typefaces require more cognitive attention to decipher at the most basic level (Keage et al., 2014), increased visual complexity during reading tasks is likely to disrupt the ability to respond to instruction with speed and accuracy. It is, therefore, important to consider what impact the visual complexity of typefaces may place on cognitive processing tasks, like automatic reading.

The Stroop Task

We measured reaction times across typefaces varying in complexity using a standard Stroop task paradigm. The Stroop Color and Word Test is an effective experimental approach for testing a variety of cognitive phenomena, including cognitive interference and automatic processing (Brown et al., 2002; Hanslmayr et al., MacLeod, 1991; Stroop, 1935). A standard Stroop task often involves presenting participants with lists of words that name colors, which are presented in either a congruent ("brown" printed in brown color) or incongruent ("brown" printed in blue color) text color, demonstrated in Figure 2. The task typically involves two tests: a "name the color" test, where participants must identify the color of the text and ignore the meaning of the word; and a "name the word" reading test, where participants must read the word and ignore the text color.

Figure 2

Congruent stimuli is consistent across the word and the print colour; whereas, with incongruent stimuli the print colour is different to the written word.

Two of the most notable findings from Stroop task research are interference and asymmetry. Stroop interference is characterized by incongruent stimuli producing slower reaction times (RT) compared to congruent stimuli, and is proposed to arise from the conflicting semantic representation of the incongruent color and text (Dalrymple-Alford, 1972; Klein, 1964; Roelofs, 2003). Stroop asymmetry describes a more pronounced interference pattern for the color naming test when compared to the word reading test. For example, in the color naming test, incongruent stimuli generate considerably slower RTs than congruent stimuli, whereas the difference in RTs between incongruent and congruent stimuli in the word reading test is less prominent (MacLeod, 1991; Stroop, 1935). This asymmetry is understood to arise from stronger automatic processing in reading compared with color identification. Given word meaning is obtained faster and without active attention; this results in a greater presence of conflicting semantic representations in the color naming test compared with the word reading test (MacLeod, 1991).

Hypotheses

This experiment is concerned with the impact of typeface complexity, defined by exaggerated stroke length, on automatic reading processes. The findings promise to not only contribute to theories of word recognition, but may also be used to improve functional readability by optimizing reading speed and comprehension and providing a better understanding of the role visual complexity plays in reading fluency. These are important factors for both font and text design. We hypothesized (H1) that we would replicate previous Stroop task findings by observing an interference effect, demonstrated by slower RTs for incongruent stimuli compared with congruent stimuli. We expected (H2) that we would also observe interference asymmetry, demonstrated by a larger interference effect when participants are asked to name the color the word is printed in (name-color test) compared with naming the word that is written out (name-word test). Further, we expected to replicate the disfluency effect (H3) by observing slower RTs with increasing typeface complexity for the name-word test. It was not expected that this would be observed in the name-color test because typeface complexity should not reduce the ability of participants to identify print color. Lastly, we expected (H4) that there would be differences in the pattern of RTs when typeface complexity, congruency, and test are considered, and that RTs for incongruent stimuli in the name-color test decrease with increasing typeface complexity (which would be the opposite pattern of the name-word test). This is because the disfluency effect should interfere with automatic word processing, thus reducing the capacity of conflicting semantic representations to inhibit text color identification (i.e., reducing the interference effect). RTs for congruent stimuli should not differ as a function of typeface complexity in the color naming test, as interference is not present.

Experiment

We measured reaction times (RTs) using an online standard Stroop task paradigm of word reading (name-word) and color naming (name-color) across four font stimuli gradually increasing in stroke length.

Participants

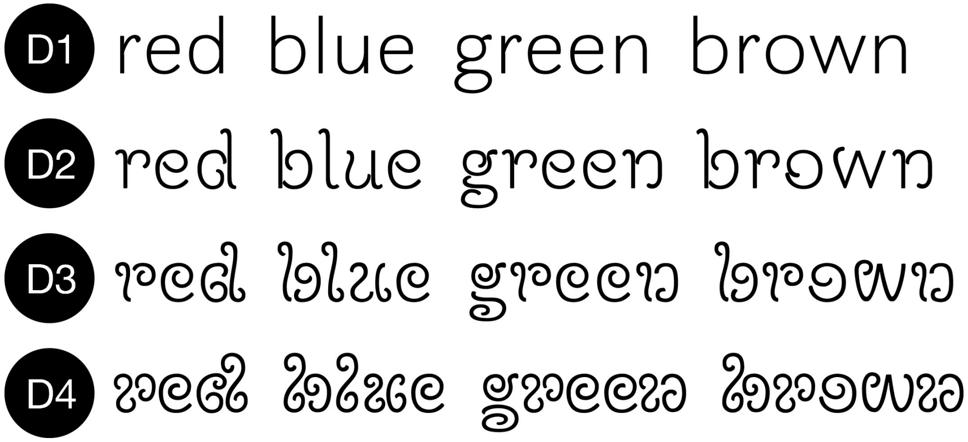
Participants were recruited using the online recruitment platform Prolific (prolific.co), and were paid a competitive honorarium. Approval was obtained from Monash University Human Ethics Low Risk Review Committee. Participants were required to read a participant information statement before beginning the experiment and consented to take part by clicking into the task window and completing the task. Participants were recruited from all countries, but were required to have completed or be currently enrolled in a Bachelor's degree program ensuring skilled reading capabilities. All participants self-reported being fluent in English, having normal or corrected to normal vision, and normal color processing. Data from a total of 200 participants were included in the analysis. There were 98 female and 102 male participants, and the average age was 23.6 years.

Materials

Single-word stimuli describing each of the four colors (red, blue, green, and brown) were presented in lower case letters in one of the four test typefaces at an x-height[3] of 40px and appeared centrally on a white background at a resolution of 150ppi. Building on Beier et al. (2017), the test typefaces are a variable font format designed for use in this experiment. The family consists of four typeface variations developed from a consistent letter skeleton, allowing us to isolate the stroke length. Shown in Figure 3, the four test typefaces can be located on a scale with the NeutralTest 1 at one extreme, following ideas of a universal letter skeleton (Frutiger, 2008), and NeutralTest 4 at the other extreme, being highly complex with stroke exaggeration that distorts the basic letter skeleton. The two remaining typefaces (NeutralTest 2 and 3) were interpolated between the outer extremes. The perimetric complexity of each typeface variation was measured and shown in Table 1.

Figure 3

Drawing on the test fonts of Beier et al. (2017), the test fonts gradually increase in length of the stroke across four levels of increased swash embellishments.



---

[3] The x-height is a variable measurement determined by the height of the lower-case letter x.

TABLE 1

Summary of the perimetric complexity of each letter used across the 4 tested fonts.

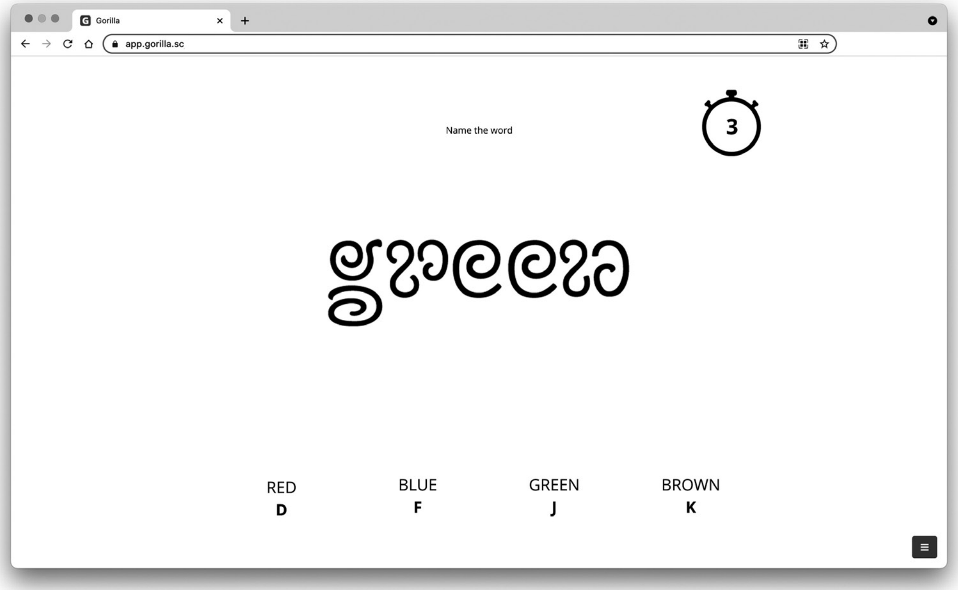
	Font			
Letter	**D1**	**D2**	**D3**	**D4**
b	114.06	142.72	165.76	203.58
d	117.17	151.11	181.99	221.75
e	102.42	100.5	108.28	110.68
g	152.5	188.22	200.82	211.91
l	57.61	66.51	89.92	113.13
n	92.71	107.46	130.12	155.77
o	90.56	98.97	106.65	117.77
r	54.53	78.03	113.09	138.19
u	89.1	107.11	120.36	157.45
w	129.12	130.83	149.78	151.09
Average	99.978	117.146	136.677	158.132

Great care was taken to ensure that the only difference between the four test typefaces related to the stroke length. The letter skeleton, along with other typographical parameters, such as letter weight and stroke contrast, were identical across stimuli tests. One exception was the letter width, which was increased when stroke exaggeration extended into the left and right side bearings of the universal letter skeleton (this is mainly seen in the letters "b," "d," "l," "r," and "n"). Isolating typographic variables in this way is an advantage for legibility experiments since tasks that draw on a range of different font families may introduce variables that are unaccounted for in the results.

We used the Internet platform Gorilla (gorilla.sc) to administer the experiment and adapted an existing Stroop template. An example of the stimulus presentation is shown in Figure 4. We included a guide at the bottom of each screen to support participants in correctly selecting the corresponding keyboard letter, and eliminate the likelihood of errors associated with incorrect recall. The colors and corresponding keys remained consistent across the entire experiment and were selected for their proximity on keyboards and typical finger placement for typing. We were also conscious that including additional text-based information should look as different as possible from the stimuli text, and all instructions were presented in a default sans serif typeface determined by each participant's browser settings.

Figure 4

Example of stimulus. We adapted a Gorilla Stroop template by including reference to the key colour mapping and since participates competed both the name-word and name-colour tests in a single sitting we included instruction with each stimulus to reduce the likelihood of errors based on confusion.



Procedure

Novel to legibility research, we drew on the online platform Gorilla to host and administer the experiment, which meant that participants completed the experiment using their own devices. The advantage of this approach is that participants are familiar and comfortable with their devices and how those are set up, and are therefore more likely to be able to use the devices with proficiency. Participants were also able to complete the study at a time that suited them. We cannot know the personal setup of each participant, but we were able to specify that the study was completed on a desktop or laptop, as opposed to a tablet or mobile phone.

The study took approximately 20 minutes to complete. Participants were shown stimuli across two naming tests: (1) name-word, in which participants were asked to ignore the print color and indicate the word that was written out, and (2) name-color, where they were asked to name the color the word appeared in and ignore the word that was written out. Both tests were presented in a congruent stimulus (e.g., color blue in the written word "blue") and an incongruent stimulus (e.g., color blue in the written word "brown"). Participants were shown stimuli in six blocks of 64 stimuli (25% congruent, 75% incongruent) where each word stimulus across the four test typefaces was presented in each of the four corresponding colors (red, blue, green, and brown). Stimuli were presented one at a time in random order for up to 3 seconds and participants responded by pressing a key on their keyboard corresponding to the colors. Participants were required to successfully complete practice rounds before each naming condition to 90% accuracy, up to three rounds. Only after the successful completion of practice were participants able to progress to the main study. This was to ensure they were familiar with the test and that they were responding quickly and with accuracy. They then completed 3 blocks for each test (name-word and name-color); the order the tests were completed in was counterbalanced across participants.

Statistical Analysis

All data processing and statistical analyses were performed using statistical packages and customized scripts on R 4.0.4 (R Core Team, 2021). Incorrect trial responses were removed (3.5% of data). Two participants were excluded due to error rates being above chance level (likely due to misunderstanding or malingering). Trials with RTs under 100ms were removed, as visual stimuli processing and motor responses physiologically cannot be enacted on these time scales. Each participant's mean error rate was then calculated for each test (i.e., name-color or name-word). Mean RTs and standard deviations (SDs) for each font disfluency level (1–4) within each test (name-colour or name-word) and stimuli congruency (congruent or incongruent) were calculated for each participant (i.e., 16 means and SDs per participant). Thirteen participants with mean RT z-scores of >3 or < –3 had their data removed to prevent extreme outliers from influencing the results. In total, 200 participants were included in our analysis.

The mean RTs were analyzed with a 2 (stimuli congruency) x 2 (test) x 4 (font complexity) repeated measures analysis of variance (ANOVA) utilizing the Greenhouse–Geisser sphericity correction method. The results of the ANOVA were considered statistically significant at $p<0.05$. A histogram of the ANOVA's residual values, in conjunction with their skew and kurtosis coefficients, were considered to ensure the data were normally distributed. Post-hoc comparisons were performed with paired sample t-tests, with statistical significance set at a Bonferroni adjusted alpha. Alphas were set at 0.025 (0.05/2) for the test x congruency interaction, 0.00417 (0.05/12) for the test x font complexity interaction, and 0.00125 (0.05/40) for the three-way interaction. Cohen's d values were calculated for each of these tests as measures of effect size.

Results

A three-way ANOVA was performed to analyze the effect of typeface complexity, test, and congruency on RT. All four sources of variance relevant to our hypotheses (congruency; test × congruency; complexity × test; complexity × test × congruency) produced significant effects, as demonstrated in Table 2.

TABLE 2

Summary of the 2 (stimulus congruency) x 2 (test) x 4 (typeface complexity) repeated measures ANOVA.

	F	η^2_G	*p*
Complexity (comparing the four typefaces)	42.48	0.012	<0.001
Test (comparing name-word and name-colour)	0.03	<0.001	0.867
Congruency (comparing congruency and incongruency)	434.35	0.045	<0.001
Complexity × Test	42.76	0.014	<0.001
Complexity × Congruency	1.21	<0.001	0.304
Test × Congruency	39.61	0.005	<0.001
Complexity × Test × Congruency	10.86	0.003	<0.001

The results replicate previous Stroop task findings, demonstrating Stroop interference and proving Hypothesis 1 (H1) in showing there was a moderate and significant main effect of congruency (see Table 2). Incongruent stimuli (M=803ms, SD=126) produced significantly slower RTs than congruent stimuli (M=748ms, SD=132).

There was a small and significant main effect of the interaction between test and congruency (see Table 2). Post-hoc comparisons revealed that incongruent stimuli (M=813ms, SD=126) produced significantly slower RTs than congruent stimuli (M=739ms, SD=128) in the name-color test. Incongruent stimuli (M=792ms, SD=126) also produced significantly slower response times than congruent stimuli (M=756ms, SD=134) in the name-word test. The size of the effect was larger for the name-color test (d=-0.58, p<0.001) than the name-word test (d=-0.29, p<0.001), showing Stroop asymmetry and supporting Hypothesis 2 (H2).

There was a small and significant effect of complexity × test interaction (see Table 2). Post-hoc comparisons revealed that there were no significant differences between the four complexity levels in the name-color test. In the name-word test, Hypothesis 3 (H3), which stated that RTs would slow as typeface complexity increased, was confirmed in that there were significant differences between the most complex stimuli (D4) and the three other complexity levels, all with moderate effect sizes (see Table 3). There were no significant differences in the comparisons between the other three complexity levels in the name-word test.

TABLE 3

Summary of the post-hoc paired sample t-tests for the typeface complexity x test interaction. Bonferroni adjusted alpha=0.00417.

Test	Complexity comparison	*p*	*Cohen's d*
Name-colour	D1 vs D2	0.010	0.10
	D1 vs D3	0.005	0.11
	D1 vs D4	0.025	-0.090
	D2 vs D3	0.84	0.008
	D2 vs D4	0.73	-0.014
	D3 vs D4	0.58	-0.022
Name-word	D1 vs D2	0.81	0.009
	D1 vs D3	0.12	-0.063
	D1 vs D4	<0.001	-0.52
	D2 vs D3	0.070	-0.07
	D2 vs D4	<0.001	-0.54
	D3 vs D4	<0.001	-0.47

There was a small and significant effect of the congruency x test x typeface complexity interaction (see Table 2 and Figure 5). For the congruent stimuli in the name-color test, a small and significant difference was observed between D2 and D4 (see Table 4), with D4 producing slower RTs than D2. For the incongruent stimuli in the name-color test, there was a significant difference between D1 and D4 (see Table 4), with D1 producing slower RTs than D4. In the name-word test, there were significant differences between D4 and the three other complexity levels for both congruent and incongruent stimuli, all with moderate effect sizes (see Table 4). There were no significant differences identified between the other complexity levels in the three-way interaction. Significant differences were observed between tests for three complexity levels (D1, D2, and D4) of the incongruent stimuli. D1 and D2 produced slower RTs in the name-color test, whereas D4 produced slower RTs in the name-word test (see Table 4). These results are consistent with the pattern that was expected under Hypothesis 4 (H4).

Figure 5

A graphical representation of the congruency x test x typeface complexity interaction. Bars indicate significant differences (Bonferroni adjusted alpha=0.00125).

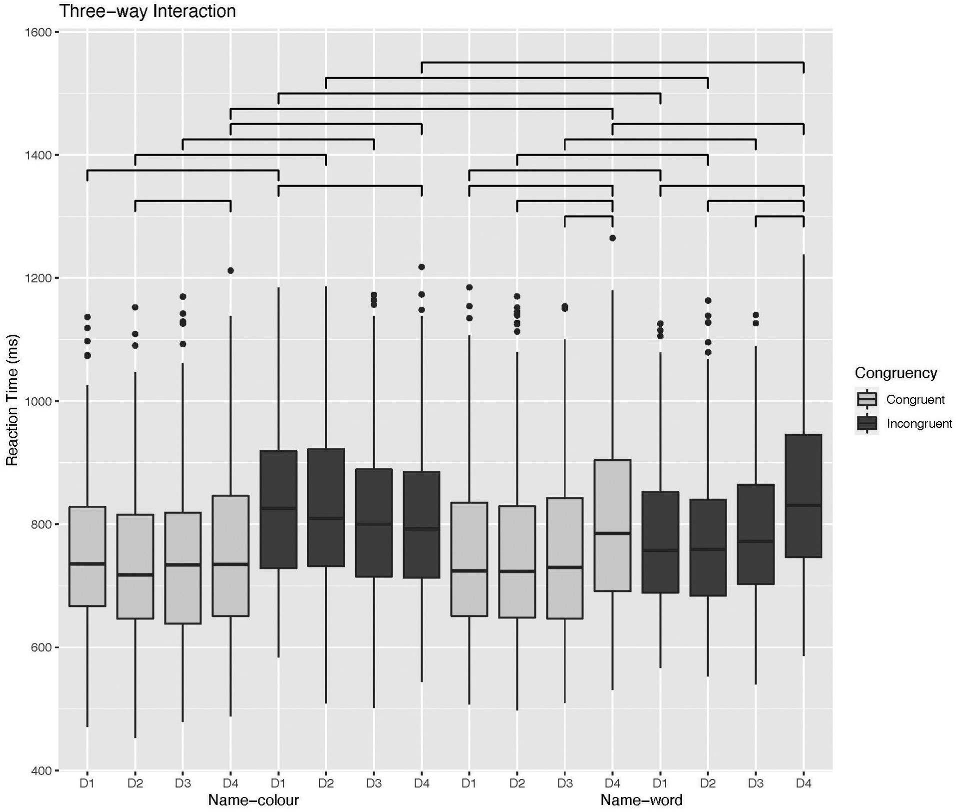


TABLE 4

Summary of the post-hoc paired sample t-tests for the congruency x test x typeface complexity interaction. Bonferroni adjusted alpha=0.00125.

Test comparisons			*p*	*Cohen's d*
Name-colour vs Name-word	Congruent	D1	0.85	0.016
		D2	0.21	-0.11
		D3	0.52	-0.049
		D4	<0.001	-0.36
	Incongruent	D1	<0.001	0.45
		D2	<0.001	0.40
		D3	0.054	0.18
		D4	<0.001	-0.40
Congruency comparisons				
Congruent vs Incongruent	Name-colour	D1	<0.001	-0.65
		D2	<0.001	-0.74
		D3	<0.001	-0.56
		D4	<0.001	-0.36
	Name-word	D1	<0.001	-0.21
		D2	<0.001	-0.23
		D3	<0.001	-0.32
		D4	<0.001	-0.40
Complexity level comparisons				
Name-colour	Congruent	D1 vs D2	0.006	0.15
		D1 vs D3	0.21	0.069
		D1 vs D4	0.35	-0.051
		D2 vs D3	0.14	-0.081
		D2 vs D4	<0.001	-0.20
		D3 vs D4	0.028	-0.12
	Incongruent	D1 vs D2	0.29	0.058
		D1 vs D3	0.005	0.16
		D1 vs D4	<0.001	0.23
		D2 vs D3	0.077	0.097
		D2 vs D4	0.002	0.17
		D3 vs D4	0.16	0.076
Name-word	Congruent	D1 vs D2	0.74	0.018
		D1 vs D3	0.89	-0.007
		D1 vs D4	<0.001	-0.43
		D2 vs D3	0.64	-0.025
		D2 vs D4	<0.001	-0.45
		D3 vs D4	<0.001	-0.43
	Incongruent	D1 vs D2	0.98	-0.001
		D1 vs D3	0.031	-0.12
		D1 vs D4	<0.001	-0.62
		D2 vs D3	0.029	-0.12
		D2 vs D4	<0.001	-0.62
		D3 vs D4	<0.001	-0.50

Discussion

Our data provide evidence for several interesting conclusions relevant for typographic research and contribute to our understanding of legibility. First, we have shown that typeface complexity, determined by stroke length, disrupts automatic reading processes; however, the fact that we only saw this effect with our most extreme typeface variation (D4) suggests that readers have a high disfluency threshold and are able to cope with high levels of typeface complexity with relative ease. Second, our data showed that the most complex typeface variation tested (D4) resulted in slower RT for the name-word test, which confirmed an expected increase in difficulty for this test that is likely the result of poor legibility (H4). In the incongruent name-color test, the opposite was the case, with faster RTs for our most complex typeface (D4) compared to the congruent stimulus. Third, as predicted, we replicated the original findings of Stroop (1935), in showing interference where incongruent stimuli resulted in slower RT (H1). We further found asymmetry between the name-color and name-word tests (H2). By matching results of other Stroop task research (Stroop, 1935), we validate the online format of the Stroop task and ensure that our new findings of word processing across multiple levels of stroke length is valid as well. This suggests value for typography and legibility research because the Stroop task can be used to index automaticity and speed, since it has shown to be a reliable measure of cognitive processing and can be used as a quick and effective tool for identifying typefaces that are likely to be problematic for readers. Further, administering the task online using platforms like Gorilla provides opportunity to test large numbers of participants quickly and efficiently.

That we saw a disfluency effect only with our most complex typeface may speak to the discrepancy in the disfluency literature. Discussed in Thiessen et al. (2020), the literature has seen considerable debate about whether difficult-to-read typefaces can improve performance with certain cognitive tasks related to memory and attention. With experiments showing inconsistent results, it is difficult to draw any definitive conclusions. In this experiment, we have shown that only extremely complex typeface variations disrupt automatic reading processes, which suggests that the lack of consensus may be attributed to whether or not the experiment stimuli were complex enough to be disruptive. Since our data have also revealed the Stroop task to be an efficient way to identify typefaces that will disrupt reading, there is opportunity to develop a better understanding about the disfluency effect with further research.

The participants who completed this experiment were university educated and skilled readers, which may account for why we did not see a lower disfluency threshold or why we did not observe RTs that correlated more directly with increasing typeface complexity (H3). Coping with and quickly tuning to a range of different typefaces and typeface styles

(Gauthier et al., 2006; Sanocki, 1987, 1988; Sanocki & Dyson, 2012) may be an important skill that is part of reading development and may be a result of exposure to a wide variety of reading materials; our participant group may be more practiced in this regard than a more diverse reading population, which may have contributed to this result. Nonetheless, our data show alignment with event-related potential (ERP) data, where Keage and colleagues (2014) analyzed ERPs following a letter recognition task and demonstrated that several stages of letter recognition were disrupted by typeface complexity. Their findings suggested that typeface complexity elicited a greater degree of perceptual attention and affected higher order cognitive processes such as visual working memory (Keage et al., 2014; Thiessen et al., 2015). This is supported by our findings of slower response times for the most complex typefaces, which likely require increased cognitive effort during reading activities.

Our most extreme typeface variation showed slower RTs for the name-word test, but interestingly, a faster RT was observed for the name-color test. This suggests that the participants may have utilized identification processes during the name-color test that were more in line with image (i.e., pictorial) than word identification. That is to say, it appears participants may not have read the words at all, but rather looked only at the display color. In effect, the typeface's complexity may have facilitated this visual image processing by reducing linguistic interference, which may be far more challenging when verbal information is presented in less visually complex typefaces and is thus more accessible. This provides further evidence that typeface complexity disrupts automatic reading processes, specifically through compromising legibility.

By developing test fonts from the same skeleton, we were able to control for other letter characteristics that might introduce unaccounted-for variables. For example, as seen in Figure 6, Times New Roman and Helvetica are familiar and often-compared fonts. These two examples differ dramatically across several stylistic features that can influence their comparison. One key difference is that Times New Roman is a serif font. In broad terms, serif fonts tend to draw influence from old style letterforms that have a long history rooted in a calligraphic tradition.[4] This tends to inform certain features, like stroke variation and the contrast between the thick and thin strokes comprising the letterform. The angle of the letter axis, serif shape, and the aperture size (Bringhurst, 1997) are also influenced. Helvetica is sans serif and demonstrates Modernist ideals that celebrate regularity and clean lines. This means that Helvetica, and fonts like it, tend to have little to no stroke variation, a vertical letter axis, and moderate to small apertures (Bringhurst, 1997). Although these are primarily stylistic features, they translate into letterforms that have substantial visual differences that may impact reading performance, such as differences in x-heights, counter space and aperture size, and ascending and descending features. These are all factors that can impact legibility and readability (Beier, 2012). By working from a single letter skeleton and isolating a single variable, we are able to say with a higher degree of certainty that any differences in performance seen in the data are related to the visual complexity resulting from an increase in stroke length. This is seen to be an important advantage for experimental designs investigating legibility; however, it is also important to recognize that the typefaces tested here have been created for testing purposes and, as a result, may lack some design features common in commercially available fonts. That is to say, they have not been designed for use in environments, and further research is needed to better understand the effect of visual complexity in realistic reading scenarios.

Figure 6

Comparing the typefaces Helvetica and Times New Roman presented at the same point size. Typefaces with a larger x-height, like Helvetica, can appear larger when compared to one with a smaller x-height, like Times New Roman, even when they are the same point size. Other stylistic differences like whether a typeface is serif or sans serif and counter and aperture size can influence legibility and

Conclusion

In a novel application of the Stroop Test, we replicated previous identified Stroop patterns (Stroop, 1935) and further showed that only a typeface of extreme complexity impaired word recognition. That is to say, on a scale ranging from a level of simple or neutral letter shapes to a level of extreme complexity—in our case, achieved by increasing letter stroke length—we found consistently slower RTs only for the extreme typeface variation when participants identified words (i.e., significantly decreased legibility) and faster RTs when they identified colors (i.e., word meaning was not interfering with color identification). These results follow multiple previous experiments that employed different experimental paradigms and collectively showed legibility impairment with visually complex typefaces (Brown et al., 2002; Hanslmayr et al., 2008; MacLeod, 1991; Stroop 1935). We add to this by demonstrating the effect solely at the extreme typeface complexity level.

---

[4] More specifically, Times New Roman is a Transitional font, which means it has some characteristics that are associated with old style letterforms, as well as other characteristics associated with more Modern serif styles that feature high contrast strokes like those seen in Bodoni, for example.

Whether or not readers are able to benefit from complex typefaces is contested, and clear and easy-to-read texts remain the desirable option. Our findings suggest that when considering reading contexts and the demands of the reading task, typefaces of great complexity should be used with caution and avoided altogether for cognitively demanding reading activities. This may be especially true when considering factors that may affect individual readers and provides an opportunity for further research. For example, specific reading difficulties like dyslexia may show differences in threshold for complexity, and children who may still be acquiring literacy skills may also show different responses. Another important consideration is the impact that environmental distractions have on attention, and when readers are expected to simultaneously process several cognitively demanding activities (e.g., driving), or when the capacity to respond to instruction is high pressure and high stake (e.g., emergency situations). Since we have shown that the Stroop task can be used to identify disfluent typefaces, further research examining how typeface complexity impacts functional reading activities, such as learning tasks or following instructions, as well as during more high-stakes scenarios, may be undertaken more effectively.

Acknowledgement

The authors thank Adam Lenzinger for contributing to the design of the test fonts, and Jean-Baptiste Bernard for providing the Perimetric Complexity calculation for this experiment.

References

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

Beier, S., Sand, K., & Starrfelt, R. (2017). Legibility implications of embellished display typefaces. *Visible Language, 51*(1), 112.

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

Robert, B. (1997). *The elements of typographic style*. Hartley & Marks: Vancouver.

Brown, T. L., Gore, C. L., & Carr, T. H. (2002). Visual attention and word recognition in Stroop color naming: Is word recognition "automatic?" *Journal of Experimental Psychology: General, 131*(2), 220.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review, 108*(1), 204.

Dalrymple-Alford, E. C. (1972). Sound similarity and color-word interference in the Stroop task. *Psychonomic Science, 28*(4), 209–210.

Diemand-Yauman, C., Oppenheimer, D., & Vaughan, E. (2011). Fortune favors the bold (and the italicized): effects of disfluency on educational outcomes. *Cognition, 118*, 111–115.

Dyson, M. & Brezina, D. (2021). Exploring disfluency: Are designers too sensitive to harder-to-read typefaces? [online], available from: https://designregression.com/research/exploring-disfluency-are-designers-too-sensitive-to-harder-to-read-typefaces, [Accessed 26 September 2022].

Fiset, D., Blais, C., Ethier-Majcher, C., Arguin, M., Bub, D., & Gosselin, F. (2008). Features for identification of uppercase and lowercase letters. *Psychological Science, 19*(11), 1161–1168.

Frutiger, A. (1989). *Signs and symbols: Their design and meaning.* New York: van Nostrand Reinholt.

Frutiger, A. (2008) The history of linear, sans serif typefaces [online], available from: http://www.linotype.com/2258/introduction.html, [Accessed 26 September 2022].

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.

Geller, J., Davis, S. D., & Peterson, D. J. (2020). Sans forgetica is not desirable for learning. *Memory, 28*(8), 957–967. doi:10.1080/09658211.2020.1797096

Grainger, J., Rey, A., & Dufau, S. (2008). Letter perception: From pixels to pandemonium. *Trends in cognitive sciences, 12*(10), 381–387.

Hanslmayr, S., Pastötter, B., Bäuml, K. H., Gruber, S., Wimber, M., & Klimesch, W. (2008). The electrophysiological dynamics of interference during the Stroop task. *Journal of cognitive neuroscience, 20*(2), 215–225.

Keage, H., Coussens, S., Kohler, M., Thiessen, M., & Churches, O. (2014). Investigating letter recognition in the brain by varying typeface: an event-related potential study. *Brain and Cognition, 88,* 83–89.

Klein, G. S. (1964). Semantic power measured through the interference of words with color-naming. *The American Journal of Psychology, 77*(4), 576–588.

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.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin, 109*(2), 163.

Oppenheimer, D. M. (2008). The secret life of fluency. *Trends in cognitive sciences, 12*(6), 237–241. doi:10.1016/j.tics.2008.02.014

Pelli, D. G., Burns, C. W., Farell, B., & Moore-Page, D. C. (2006). Feature detection and letter identification. *Vision Research, 46*(28), 4646–4674.

Petit, J.-P., & Grainger, J. (2002). Masked partial priming of letter perception. *Visual Cognition, 9*(3), 337–353.

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.

Reichle, E. D. (2020). *Computational models of reading: A handbook*. Oxford University Press.

Roelofs, A. (2003). Goal-referenced selection of verbal action: Modeling attentional control in the Stroop task. *Psychological Review, 110*(1), 88.

Sanocki, T. (1987). Visual knowledge underlying letter perception: Font-specific, schematic tuning. *Journal of Experimental Psychology: Human Perception and Performance, 13*(2), 267.

Sanocki, T. (1988). Font regularity constraints on the process of letter recognition. *Journal of Experimental Psychology: Human Perception and Performance, 14*(3), 472.

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.

Sawyer, B. D., Wolfe, B., Dobres, J., Chahine, N., Mehler, B., & Reimer, B. (2020). Glanceable, legible typography over complex backgrounds. *Ergonomics, 63*(7), 864–883.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*(6), 634–662.

Taylor, A., Sanson, M., Burnell, R., Wade, K. A., & Garry, M. (2020). Disfluent difficulties are not desirable difficulties: The (lack of) effect of Sans Forgetica on memory. *Memory*, 1–8. doi:10.1080/09658211.2020.1758726

Thiessen, M., Beier, S., & Keage, H. (2020). A review of the cognitive effects of disfluent typography on functional reading. *The Design Journal, 23*(5), 797–815. doi:10.1080/14606925.2020.1810434

Thiessen, M., Kohler, M., Churches, O., Coussens, S., & Keage, H. (2015). Brainy type: A look at how the brain processes typographic information. *Visible Language, 49*(1/2), 175–189.

Walczyk, J. (2000). The interplay between automatic and control processes in reading. *Reading Research Quarterly*, *35*(4), 554–566.

Authors

Myra Thiessen*
Art, Design, and Architecture, Monash University,
Melbourne, Australia
myra.thiessen@monash.edu
0000-0003-2887-2129

Myra Lecturers in Communication Design in the Department of Design and is a researcher in the Design Health Collab at Monash University. Her research is focused on design for reading with a particular interest in how motivation, context, and environment affect comprehension and decision making in healthcare settings.

Hannah Keage
Cognitive Ageing and Impairment Neurosciences
Laboratory, UniSA Justice & Society, University of
South Australia, Adelaide, Australia
hannah.keage@unisa.edu.au

Hannah obtained her PhD from Flinders University, in South Australia. She undertook post-doctoral positions at the University of Cambridge between 2007 and 2011, before taking up an academic position at the University of South Australia (UniSA). She is currently an Associate Professor of Psychology at UniSA.

Indae Hwang
Art, Design, and Architecture, Monash University,
Melbourne, Australia
indae.hwang@monash.edu

Indae is a Melbourne-based interactive artist and designer, researcher and lecturer in the Department of Design at Monash University. His teaching focuses on exploring new ways of embracing and utilising emerging media technologies in the context of User Experience and Interactive Design.

Jack Astley
Cognitive Ageing and Impairment Neurosciences
Laboratory, UniSA Justice & Society, University of
South Australia, Adelaide, Australia
jack.astley@unisa.edu.au

Jack completed his undergraduate and Honours degree in Psychological Science at the University of Adelaide. He is currently employed as a research assistant for the Cognitive Ageing and Impairment Neurosciences Laboratory at the University of South Australia, in addition to working as Data Analyst at Inventium, a behavioural science consultancy.

Sofie Beier
Centre for Visibility Design, Royal Danish Academy:
Architecture, Design, Conservation, Copenhagen,
Denmark
sbe@kglakademi.dk

Graphic designer and professor WSR, Sofie is employed at the Royal Danish Academy, where she is head of Centre for Visibility Design. She is the author of the 'Type Tricks' book series and of 'Reading Letters: designing for legibility'. Her research is focused on improving the reading experience by achieving a better understanding of how different typefaces and letter shapes can influence the way we read.