Perception of Letter Glyph Parameters for InfoTypography

JOHANNES LANG, Languste Fonts and University of Applied Arts Vienna, Austria
MIGUEL A. NACENTA, University of Victoria, Canada

weight	width	contrast	xHeight	slant	serifs	aperture
infotypography	infotypography	infotypography	infotypography	infotypography	infotypography	infotypography
infotypography	infotypography	infotypography	infotypography	infotypography	infotypography	infotypography
infotypography	infotypography	infotypography	infotypography	infotypography	infotypography	infotypography
infotypography	infotypography	infotypography	infotypography	infotypography	infotypography	infotypography
infotypography	infotypography	infotypography	infotypography	infotypography	infotypography	infotypography
infotypography	infotypography	infotypography	infotypography	*infotypography*	infotypography	infotypography
infotypography	infotypography	infotypography	infotypography	*infotypography*	infotypography	infotypography
infotypography	infotypography	infotypography	infotypography	*infotypography*	infotypography	infotypography
infotypography	infotypography	infotypography	infotypography	*infotypography*	infotypography	infotypography

Fig. 1. The word "infotypography" rendered in variations of seven type parameters: weight, width, contrast, xHeight, slant, serifs and aperture.

The advent of variable font technologies—where typographic parameters such as weight, x-height and slant are easily adjusted across a range—enables encoding ordinal, interval or ratio data into text that is still readable. This is potentially valuable to represent additional information in text labels in visualizations (e.g., font weight can indicate city size in a geographical visualization) or in text itself (e.g., the intended reading speed of a sentence can be encoded with the font width). However, we do not know how different parameters, which are complex variations of shape, are perceived by the human visual system. Without this information it is difficult to select appropriate parameters and mapping functions that maximize perception of differences within the parameter range. We provide an empirical characterization of seven typographical parameters of Latin fonts in terms of absolute perception and just noticeable differences (JNDs) to help visualization designers to choose typographic parameters for visualizations that contain text, as well as support typographers and type designers when selecting which levels of these parameters to implement to achieve differentiability between normal text, emphasized text and different headings.

CCS Concepts: • **Computing methodologies → Perception**; • **Human-centered computing → Information visualization**.

Additional Key Words and Phrases: Typography, Visualization, Infotypography, Glyphs, Type Design, Variable Fonts, Document Design.

ACM Reference Format:
Johannes Lang and Miguel A. Nacenta. 2022. Perception of Letter Glyph Parameters for InfoTypography. *ACM Trans. Graph.* 41, 4, Article 147 (July 2022), 22 pages. https://doi.org/10.1145/3528223.3530111

Authors' addresses: Johannes Lang, Languste Fonts and University of Applied Arts Vienna, Vienna, Austria, jo@langustefonts.com; Miguel A. Nacenta, nacenta@uvic.ca, University of Victoria, Victoria, Canada.

1 INTRODUCTION

Font designs vary in many significant ways; for example, some fonts will render text with thicker characters (i.e., heavier weights), others have narrower letters (narrow width), or more prominent serifs. These parameters constitute part of the design space of fonts, and type designers use different weights, widths, contrast or slant to create designs that are unique, to convey particular messages (e.g., a font might feel classic or contemporary, serious or playful), or to support different purposes in the text (e.g., heavier weights—bold text—highlight or emphasize parts of text in a document).

With the popularization of variable font technologies, also called parametric fonts (e.g., [Hudson 2016; Sherman 2021]), it has become much easier to design, implement, and use fonts that vary continuously in one or more parameters. Whereas type designers would traditionally design a discrete choice of weights, slants and widths for a family of fonts (e.g., semi-light, heavy, bold, narrow, ultra-wide), variable fonts allow the user to select virtually any values of the typographic parameters that serve their specific purpose. For example, a poster designer can choose the weight and width that will make a title fit a specific page size with adequate prominence.

Several researchers have suggested that the continuous typographic parameters offered by variable fonts can be used, instead, to convey quantitative information. For example, streets in a city map that are busier can be rendered in heavier weights proportional to their traffic throughput [Afzal et al. 2012] and syllables in song lyrics can be rendered with different heights depending on the corresponding pitch [Brath and Banissi 2016]. We call this use of type design parameters *InfoTypography*. In this work we are chiefly concerned with the use of letter shape parameters in parametric fonts to represent ordinal, interval, or ratio scalar values.

Although infotypographic techniques are promising for visualization, little empirical evidence exists on how humans perceive

typographic parameters. This is important because if different parameters are subject to different levels of perceptual noise and bias (i.e., different levels of error when reading the information encodings) then one should be able to chose the most effective ones.

Here we present two experiments characterizing human perception of seven typographical parameters in Latin fonts: weight, width, contrast, x-height, slant, serifs and aperture. Both studies took place over the Internet. In the first study, 214 participants estimated the level of a typographical parameter at which a word was rendered. In the second, 237 participants compared words rendered at different levels of a parameter in a staircase procedure to determine the just noticeable differences (JNDs) along each parameter's range.

The results rank the parameters in increasing order of perceptual noise, i.e., in terms of the variability in estimations of the level of a stimulus. Weight and slant are perceived most accurately, followed by xHeight and width, then contrast and serif size, with aperture—which is combined with junction—as least accurately perceived. The results facilitate the selection of parameters for data representation. Additionally, we characterize each parameter's perception curve. Compensating for these curves enables a more efficient distribution of parameter points for mapping ordinal levels or supports perceptual equalization of parameters for continuous values.

This research can help several groups of professionals and end users of Latin fonts: visualization designers can use our results to choose appropriate typographic parameters and improve the mappings between parameters and continuous or categorical data; type designers can complement their visual instincts and expertise with our models to design fonts with typographic parameters at more perceptually uniform intervals; typographers, document template designers and UI designers can choose visually distinct font variants for different functions in the page or screen based on human ability to differentiate font variations; and, researchers of type and reading can leverage our models to explain the effects of typographic variations on reading and perception of type.

2 BACKGROUND AND RELATED WORK

Here we introduce the concepts underlying type design, typography, and data visualization and highlight relevant previous research.

2.1 Reading and Type

Letter glyphs are visual representations of the characters of a script system.[1] The core purpose of glyphs is to allow visual recognition of the characters (letters) so that reading can take place. Letter recognition by most literate people's visual system is effective [Dehaene 2010; Rayner et al. 2012]; a large variety of glyph shapes are successfully and efficiently recognized as the same letter (e.g., a, *a*, **a**, **a**, a) [Hofstadter and Sander 2013, p.5]. This flexibility enables a broad diversity of typeface designs, most of which can still be read effortlessly by most people.

The large possible variation in glyph shapes and the spatial relationships between those glyphs (e.g., kerning) is the domain of type designers, whose job is to design glyphs that support reading (readability and legibility) but also that convey other characteristics of the text such as formality, or that provide aesthetic value and capture attention (e.g., for advertising). The shape of glyphs can sometimes also provide historical context for the text because specific styles are associated to certain periods [Bringhurst 2004; Hyndman 2016]. Additionally, fonts can convey emotion [Choi et al. 2016]. Glyph shape variation is therefore able to encode information beyond the representation of the alphabet's characters.

Typographers, who are responsible for the visual appearance of printed matter, use type selection as a key design tool to compose practical and pleasant layouts on a variety of formats (including paper and digital). For example, a well-designed corporate report will convey its structure effectively through headings in a selection of different sizes and font variants that make it easy to grasp the structure and content of the report while still looking professional.

2.2 Studies on Fonts

There is a long history in the study of how typefaces and typographical parameters (e.g., serifs) affect readability and legibility (e.g., [Arditi and Cho 2005; Beier et al. 2017; Chaparro et al. 2010; Dobres et al. 2016; Mansfield et al. 1996; Tinker 1944]), and of how to design typefaces to maximize legibility and readability (e.g., [Beier 2012]). Recently, Kadner et al. produced Adaptifont [2021], a system to increase people's reading speed through a variable font and a machine-learning-driven adaptive procedure. Although considerations of readability, legibility and reading speeds are important for font design and use, here we focus on the differentiability of fonts based on the variation of typographic parameters.

2.3 Infotypography

We define InfoTypography as *the use of variations of letter shapes and their spatial arrangement within text to visually represent quantitative or qualitative information beyond the textual content.* Strictly speaking, InfoTypography is almost as old as writing itself because variations in script size and spacing have long been used to visually differentiate parts of the text such as headlines or footnotes, according to their function (an instance of categorical information). An example of this approach is Strobelt et al.'s study of typographic highlighting [2016], which tested how nine typographic parameters (e.g., italics, bold) and non-typographic parameters (e.g., font color, rectangular border) facilitate finding highlighted text. They found that font size provided the most effective highlighting (individually).

Map makers have long used type size and other variations, such as italics, to demarcate categories of named items in maps [MacEachren 2004, p.128,293]; for example, text size in a political map is often matched, in decreasing order, to continents, country names, and then cities (sometimes with variation depending on the city's population). Maps also typically use italics for rivers or mountain ranges. This highlights InfoTypography's value as a technique to overload information where additional visual objects or text interfere with or cause overcrowding of a graphic. The discriminability of type size in map contexts was studied by Shortridge [1979]; a follow up study considered also capital letters, boldness, and size, indicating that multiple encoding leads to better discrimination up to two parameters (three parameters in particular cases) [Shortridge and Welch 1982]. A contemporary infotypographic use in maps are Afzal et

[1] This working definition is still controversial [Haralambous and Haralambous 2003].

al.'s maps [2012], which explicitly introduce the idea of encoding continuous values in the text size or weight of a street's name.

More abstract uses of font parameters are relatively common in visualization, with tag clouds (a.k.a. word clouds) being probably the most popular. Bateman et al. [2008] measured the perceived salience of tags based on nine visual features, including color, saturation, font size and boldness (bold and normal), and found that size, weight and color are the visual characteristics that most strongly determine subjective salience. Several others have studied tag clouds, but mostly with an emphasis on non-typographic aspects of the representation (e.g., [Alexander et al. 2018; Felix et al. 2018]). Another example of font manipulation for visualization are FatFonts: digits that increase weight to double-encode numerical values [Han and Nacenta 2020; Manteau et al. 2017; Nacenta et al. 2012].

Over a series of articles, Brath and Banissi provided a comprehensive examination of current and potential infotypographic applications in visualization and beyond [2014a; 2014b; 2016; 2017]. They also point out how Bertin's seminal work on the Semiology of Graphics [1967] included a section on typographical variables that was not included in the 1983 English translation [Brath and Banissi 2019]. Brath and Banissi make a compelling argument for the utility of typographic variables. It is plausible that InfoTypography will become more widely useful and sophisticated due to the increasing availability of technologies to design and render variable fonts.

2.4 Variable and Parametric Typography

The first major proponent of the use of software to generalize type design was Donald Knuth with METAFONT [1986], a language and interpreter that enables the design of glyphs and collections of glyphs through declarative high-level constraints that can then be adjusted according to 62 parameters such as weight and xHeight. Since Knuth's late 1970's and early 1980's work, several commercial systems have allowed automatic generation of type families, mostly with the goal of facilitating creation of font families with multiple weights (e.g., Infinifont [McQueen III and Beausoleil 1993]). Probably the most successful was Adobe Multiple Master technology, which allowed the interpolation between compatible outlines of a glyph [Adobe Corporation 1997]. Eventually, Multiple Master technology was superseded by OpenType, a current major font format [Microsoft 2018]. In parallel, Apple has long used their own extension of TrueType font technology (Apple Advanced Type Technology—AAT [Apple Corporation 2020]), which allows for sophisticated typographical variations in one or more *styles* (a similar concept to the meta-parameters of METAFONT).

Most recently, Adobe, Apple, Google and Microsoft extended the OpenType standard into OpenType Variable Fonts [Microsoft 2018]. The standard allows the combination of several styles (parameters like weight or width) into one font file (a variable font), which can be rendered through interpolation into a very large number of variants. This generalizes previous independent work such as prototypo.io's customized type designer [Mathey and Prototypo Team 2020], which we used as base rendering engine for our study. Some herald OpenType Variable Fonts as the future of typographic formats in the web [Hudson 2016; Nieskens 2016]. Variable Fonts are already widely supported by the major browsers. Importantly, the manipulability of the font's parameters requires significant human effort: which parameters are available and how they vary (e.g., whether the variation is linear or otherwise with respect to the parameter) is still to be designed by the type designer and the implementer of the font. Most fonts implemented so far only vary in one to three styles (parameters—see Ref. [Sherman 2021] for a compilation). It is therefore timely and important to have an empirically-supported understanding of the perceptual characteristics of at least the most important parameters for typographical display as well as for infotypographic uses of this technology.

In parallel, a history of academic efforts has also supported the design and implementation of variable fonts. For example, Haralambous [1993] turns PostScript fonts into parametric fonts through METAFONT, Shamir et al. [1998] propose a system for parametric fonts designed by humans aided by hierarchical constraints, and Hu and Hersch [2004] built a system using parametric characters to generate perceptually-tuned raster renderings of fonts. Hassan et al. [2010] provide a summary of previous efforts, and Brownie [2007] theorizes about the implications of typefaces that can change.

More recently, machine learning and computer vision spurred a revival in computational type research. Advances support cataloguing and understanding the existing landscape of available fonts (e.g., [Shinahara et al. 2019; Uchida et al. 2015]), but most importantly, generating, transferring or interpolating typefaces for different purposes, including variable fonts [Atarsaikhan et al. 2017; Azadi et al. 2018; Balashova et al. 2019; Baluja 2017; Campbell and Kautz 2014; Gao et al. 2019; Hayashi et al. 2019; Kadner et al. 2021; Lian et al. 2018; Miyazaki et al. 2020; Phan et al. 2015; Wang et al. 2020]. These advances will likely facilitate the design of variable fonts in new ways and for many purposes, including for InfoTypography.

3 A PARAMETRIC FONT FOR INFOTYPOGRAPHY

As a base for our investigation we modified the *John Fell* parametric font used by the *Prototypo.io* service.[2] The original font is written in Javascript and takes many parameters such as weight, width, serif-length, serif-width, roundness, slant and descender length. The parameters support the design of font styles according to aesthetic choice; unfortunately, they are not directly appropriate for the representation of information for four main reasons: a) some parameters do not vary in a predictable way, b) some parameters affect a small subset of the letters (e.g., descenders), c) some parameters are too subtle to perceive at first glance, and d) some parameters interact with other parameters too much, making the overall font appearance unstable, the text unreadable, or some letters indistinguishable from each other. Our modified version merges some of these parameters, makes others more appropriate for InfoTypography, and leaves the rest at their default values.

We sought to select a set of parameters for study that met the following criteria: a) the list should include most of the common variations from the type design tradition; b) the parameters should be intrinsic to letter shape; c) the parameters should support continuous variation; d) the set should include the most common variations provided by existing vfariable fonts; and e) it should be possible

[2] PTFell font, see http://www.prototypo.io.

to vary all parameters in the set simultaneously. We chose these criteria to maximize the relevance of the study to practitioners in variable and traditional type design, typographers, and visualization designers (further detail in Appendix A.1).

The set of controllable parameters has: weight, width, contrast, xHeight, slant, serifs (serif-length), and aperture combined with junction angle (see Figures 1 and 2). In general those parameters correspond to their meaning in typographic terms [Beier 2012; Cheng 2020; Coles 2013; Haralambous 2007; Lupton 2010], except of aperture. For example, the weight parameter controls the thickness of the shapes linearly, with vertical strokes as reference. Our xHeight defines the height of the main body of the letter as a proportion of a fixed vertical height that includes ascenders from the baseline. Therefore a larger xHeight results in more slender glyphs that also have shorter ascenders. Aperture is an exception because it is a conglomeration of aperture (the space between finial-terminated openings; e.g., between the counter and the tip of the bottom stroke of an 'e') and junction, which defines the depth at which the shoulder of the 'n' or the bowl of the 'd' join their respective stems. We decided to combine the two because we anticipated that either of them would be too subtle individually and because together they are more likely to affect the shape of a word (neither affects all letters).

The ranges of the parameters in our font and experiment are displayed in Figure 2 and the geometrical measurements that they correspond to are shown in Table 1. All measures are in *units per em* (UPM),[3] except for slant, which uses degrees. Ranges were chosen iteratively according to two criteria: a) we wanted the values to generate reasonably practical and readable fonts (e.g., text with letters more than twice as wide as tall become noticeably awkward to read and to use in paragraphs); and, b) we needed to avoid artefacts in the letters due to their implementation. The final product is an improved parametric font for InfoTypography and a set of usable ranges and neutral points that define a hypercube in 7 dimensions for which we are reasonably sure that any combination of parameters results in a readable font instantiation.[4] Our infotypographic font provides much broader coverage than typical parameter ranges. For example, italic or slanted versions of fonts rarely go beyond 12° angles (our range is -20° to 20°). Although a few fonts have more extreme levels, these are typically not usable for readable text.

Table 1. Geometry and range of typographical variables. 'Ref' refers to Figure 3. All units are in UPM (units per em) except for slant (measured in degrees). Aperture and junction are combined in the aperture parameter (they co-vary linearly).

Typ. par.	Geometrical description	Ref	Min	Neut	Max	Steps
weight	Stem thickness of 'n' stem	A	10	85	200	95
width	Distance between centers of 'n' stems	B	209	308	505	150
contrast	Thickness of thinnest over thickest strokes in 'o'	C	$\frac{0}{113}$	$\frac{36}{113}$	$\frac{71}{113}$	40
xHeight	Height of the 'x'	D	300	500	600	100
slant	Angle of vertical stems	E	−20°	0°	20°	80
serifs	Width of base serifs in 'm'	F	0	65	85	85
aperture	Size of opening in the 'e'	G	64	135	180	52
junction	Vertical distance b/w stem joint and hump of 'n'	H	132	82	50	52

4 RESEARCH GOALS AND QUESTIONS

The overarching goal is to measure people's perceptual ability to distinguish variations in glyph parameters to inform infotypographic applications. For this purpose, we seek to achieve four main objectives: first, to characterize perception of glyph parameters in order to know which ones are less noisy as visual channels to represent information; second, to quantify the accuracy of human observers at different values of these parameters so that designers of infotypographic applications can better decide how to map values optimally (instead of assuming a linear scale); third, to learn which levels of the parameters form perceptually uniform intervals so that type designers can decide which levels in a multi-parameter font to implement (i.e., to allow typographic designers to spend more time and effort in the font variants that are most likely to be perceived as different from other font variants); and fourth, to support typographers or designers of interfaces and document templates when choosing typographic levels that are differentiable. These objectives lead to four main questions:

- **Q1** How do the different typographic parameters compare in terms of perceptual noise?
- **Q2** How do human perceptual estimations vary with respect to the linear variation of each parameter?
- **Q3** What is the maximum number of distinguishable levels that each parameter can support?
- **Q4** What is the variability in terms of perceptual noise for different people?

To answer these questions we designed and carried out two experiments on the absolute and relative perception of glyph parameters.

5 METHODOLOGICAL APPROACH

This section describes the main methodological decisions common to both experiments. The anonymized data and the scripts for analysis are available in the Supplementary Materials.

5.1 Crowdsourcing and data pre-filtering

The experiment was crowdsourced to run a sufficient number of participants. We recruited participants for all experiments through the Prolific service[5] and redirected them to a locally run web service that presented the experiments and recorded the data. Participants were 18 or older, with a Prolific score of > 90%, live in a countries

[3] UPM is a common measure in type design, where an *em* defines the size of a square font grid that fits most glyph's height, and it is divided into a number of units (commonly 1000 units, which is what we use). See Ref. [Wikipedia contributors 2020].

[4] Static font files for the variation of each parameter are in the Supplementary Materials. To facilitate extrapolation of our results we include a table in the Supplementary Materials with our parameter measures of popular font families (e.g., Helvetica and Calibri), parametric fonts (Amstelvar), and font families with extreme values (e.g., Source).

[5] https://prolific.co

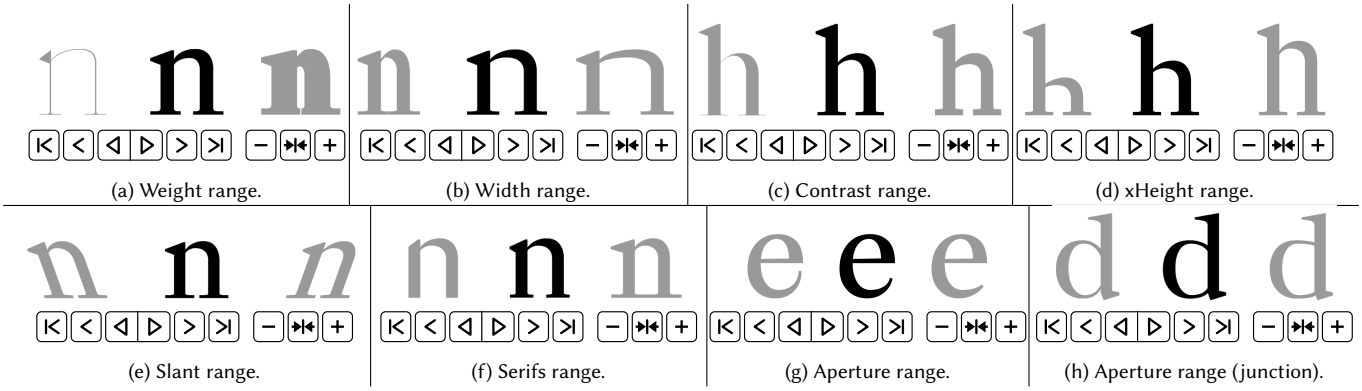

Fig. 2. Parameter ranges in our infotypographic font. The aperture parameter (which combines aperture and junction angle) is represented twice: (g) aperture and (h) junction to display the two aspects separately.

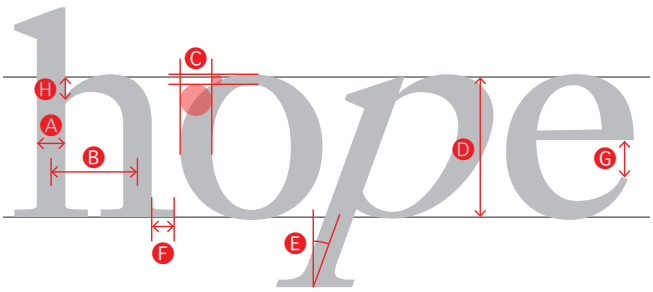

Fig. 3. Geometric bases of parameters. Letters refer to Ref column on Table 1.

where English is a dominant official language and have English as native language. Participants were compensated at minimum wage (more detail in Appendix B).

Although crowdsourcing experiments over the Internet introduces noise from variability in the circumstances and appearance of the experiment, it also provides a degree of ecological validity [Heer and Bostock 2010; Woods et al. 2015]. We ameliorated some of the variability issues by only sourcing participants classified by Prolific as using a desktop computer (most desktop computers would have resolutions adequate for our experiments). We discarded data that was not complete or deemed to be suspect (see Appendix B).

5.2 Bayesian Approach to Analysis through MCMC

We deployed a contemporary Bayesian framework for the analysis of the studies based on Markov Chain Monte Carlo (MCMC) computation of probabilities [Gelman et al. 1995; Kruschke 2014]. MCMC is the current dominant way to perform Bayesian analysis. Appendix C.1 further justifies this choice. We followed best practices following Kruschke's textbook [2014].

5.3 Range Normalization

To simplify the analysis and interpretation of results, the input and output parameters are all normalized between −0.5 and 0.5. Most values reported below are therefore dependent on our choice of ranges. Functions to normalize and denormalize the parameter ranges based on Table 1 are in the Supplementary Materials.

6 EXPERIMENT 1: ABSOLUTE PERCEPTION

This experiment aims to measure participants' absolute responses to variations in the typographic parameter of the stimuli. It assesses participants' ability to map a stimulus from one typographic parameter to a value in a linear response scale while all other typographic parameters remain at a fixed neutral value (listed in Table 1). It measures the typical error at different points in the scale, systematic deviations from the linear scale (bias), and their shape.

6.1 Task

Participants adjusted a slider (Figure 4, item 2) within a continuous scale (1) to indicate where in the scale they thought that the reference word (5) was located within the range of the scale. To signal which ends of the scale corresponded to which values, we provided example words rendered with the extreme values of the typographic parameter at the corresponding extremes of the scale (3 and 4). The words in the task screen were always different from each other. To avoid interference from word familiarity we drew all words from a list of 1197 high-frequency 8-letter words without proper nouns that we manually curated to exclude potentially disturbing words (available in the Supplementary Materials). The words were always rendered in lower-case. The slider accepted 100 different values, and participants learned before the experiment how to adjust it through their pointing device and through the arrow keys.

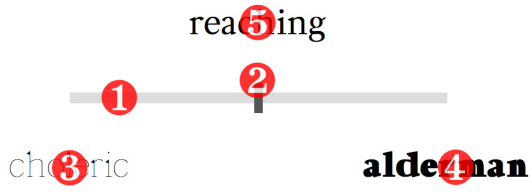

Fig. 4. Experiment 1 task as it looked for the weight parameter.

6.2 Procedure

Participants provided informed consent, then read instructions for the procedure and a detailed description of the typographic parameter being tested with graphics similar to Figure 2. They then performed 81 trials in an unpredictable and counterbalanced order (9 repetitions for each of 9 levels of the typographic parameter between -0.5 and 0.5 in the normalized parameter space—see Section 5.3).

6.3 Participants

We obtained 254 full sets of trials for this experiment of which we kept 214, a bit over 30 full datasets per parameter (19 sets which were not complete and 40 were filtered). Participants did not repeat the same parameter, but they could do multiple parameters (187 unique participants completed the 214 full sets of this experiment). To be conservative, we only carry out between-subjects analyses. 88 were female and 1 did not disclose sex. Participants were between 18 and 65 in age, with a median and mode of 30. There is further detail in Appendix D.1.

6.4 Specifics of the Analysis

The analysis aims to provide, per parameter, accurate estimations of the level of perceptual noise (Q1), a model of the relationship between stimulus and average response (Q2), a series of distinguishable stimuli levels (Q3), and an estimation of perceptual variability in the participant population (Q4). We perform model fitting in two phases: first we fit a set of models to select the best model, and then we further fit a more sophisticated version of the model that accounts for participant variability and different noise at each of the nine tested levels.

6.4.1 Model Comparison. To model the relationship between participant responses (y) to the specific parameter (stimuli) value (x), we assume a probabilistic Gaussian distribution of perceptual noise where the center of the predicted value $\mu(x)$ (the Gaussian mean) depends on the stimuli and there is a global standard deviation (σ) per model (see Eq.1).

$$P(y \mid x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(y-\mu(x))^2 \Big/ 2\sigma^2} \tag{1}$$

The shape of the bias ($\mu(x)$) can potentially be modeled by any function that maps the dimension of the stimuli (the value of the parameter) to a perceived value. Since it is not practical to compare a very large number of models, we chose four possible algebraic models to fit responses: quadratic, sigmoid, inverse-sigmoid, and cubic. The basic curves are defined in equations 2 to 5 and displayed in Figure 5.

$$\mu_q(x) = \beta_0 + \beta_1 x + \beta_2 x^2 \tag{2}$$

$$\mu_s(x) = \frac{1}{1 + exp(-\kappa_s(r_s x - x_{s0}))} - 0.5 \tag{3}$$

$$\mu_i(x) = \frac{log\left(\frac{1}{r_i x + 0.5} - 1\right)}{-\kappa_i} + y_{i0} \tag{4}$$

$$\mu_c(x) = y_{co} + a\,(x - x_{co})^3 + c(x - x_{co}) \tag{5}$$

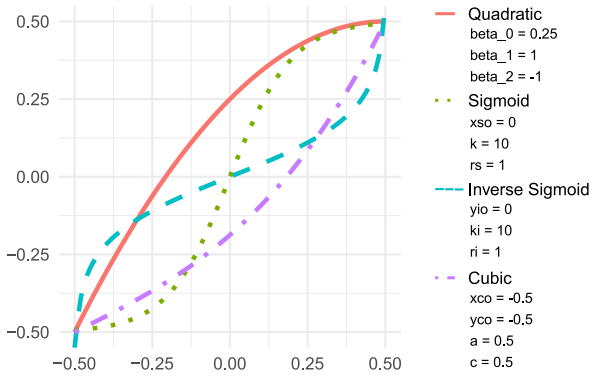

Fig. 5. The basic models (represented with illustrative parameters).

The outcome of the model comparison are Bayes Factors for each of the four model types, as well as best parameter fits for each model. The Bayes Factors represent the comparative posterior probability of a model being best at explaining the data among the models included in the comparison. We choose the model with the highest Bayes factor, i.e., the most plausible model. The Bayesian procedure takes care of regularization issues (i.e., it balances probabilities when models have more degrees of freedom—see explanation in Appendix D.2 and Ref. [Kruschke 2014, p.291]).

We chose the models to be able to reasonably approximate most possible monotonic behaviors of the curve. The models cover cases where the curve flattens as the value increases (consistent with Weber-Fechner's law [Fechner 1948]), or decreases, and curves that are more (or less) steep in a particular region of the range, which are covered by sigmoid (and inverse sigmoid) models. We did not explicitly include exponential/logarithmic curves, but these are reasonably approximated by polynomial curves. We chose parameterizations to simplify interpretation of the results. For example, the sigmoid curve has parameters that modify the vertical and horizontal offset of an s-shaped curve that crosses the center of the range, with an additional parameter that determines the slope of the curve at the center (green dotted curve in Fig 7). The representativeness of the models and rationale for parameterizations are in Appendix D.2.

To complement the probabilistic model of Equation 1, which has to use data averaged per participant, we also fit a more sophisticated model (Equation 6) that accounts for the variability in responses *per participant* by introducing a *participant multiplier* (*pMult*) of the variances. Participants that are less accurate than the average have *pMult* above 1, and more accurate participants have *pMult* below 1. Modeling the variability of *pMult* allows us to approximate the expected variation in participant perceptual accuracy across the population. This model also fits different levels of variability to different levels of the typographic parameter stimuli (i.e., σ is a function dependent on the stimulus x rather than a single variable). Note that here we only use the $\mu(x)$ function of the model of the first phase that was most likely. The multi-phase procedure and setting of prior and pseudo values to fit parameters and calculate the Bayes factors are in Appendix D.4.

$$P(y \mid x, p) = \frac{1}{\sigma(x) pMult[p]\sqrt{2\pi}} e^{-(y-\mu(x))^2 \Big/ 2(\sigma(x) pMult[p])^2} \quad (6)$$

7 EXPERIMENT 1: RESULTS

We report following research question order from Section 4. Table 2 provides an overall summary of the main results.

7.1 Q1: Perceptual Noise

Aggregate measures of standard deviation from the average on each of the nine tested levels of the stimuli appear on Table 2, columns labelled $\sigma_{R,A}$ and $\sigma_{R,U}$. These are measurements based exclusively on data (no modeling). $\sigma_{R,A}$ is smaller because it averages all the responses by a participant before calculating the standard deviation, whereas $\sigma_{R,U}$ considers individually all the individual data points from each participant. Figure 6 provides a visual summary of the results for each parameter.

The ranking of $\sigma_{R,A}$ and $\sigma_{R,U}$ places weight as the least noisy typographic parameter, followed closely by slant, with width and xHeight next (and close to each other), then contrast, and then serifs and aperture at the end (close to each other). Serifs and aperture shift places between the averaged and per-participant measurements.

When we estimate the standard deviations through fitting the best model (see Section 6.4.1) we get different values because the models account for the floor and ceiling effects of the scale through censoring (see Appendix C.3). Additionally, the per-person model enables individual multipliers of error depending on the participant's skill. The per-person estimations of standard deviation $\sigma_{M2,U}$ should be more robust estimations of noise for the different typographic parameters. These are displayed in Figure 7 and place slant as least noisy, followed by weight, width, xHeight, contrast, serifs and aperture, in that order. To provide an estimate of the reliability of that ordering, we computed the average standard deviation of responses across the nine stimulus levels with their corresponding 95% high density intervals from the MCMC chains (Figure 8). Weight and slant show a small overlap and width and xHeight show a more significant overlap, with all other intervals being clearly distinct.

7.2 Q2: Curve Shapes

The Bayesian model comparison selects the most likely shape of the perceptual responses to the stimuli (Table 2, column 6).[6] We were surprised by the variety of models that were best fits for the data from different typographic parameters. Data from weight and width were fitted best with quadratic models with negative coefficients of the square term (a concave shape flattening as the parameter value increases). Contrast is cubic, but with a concave shape, approximately the inverse behavior of weight and width. For xHeight, serifs and aperture the best fit was a sigmoid curve, where the slope is steepest roughly in the middle of the range, and flattens when it approaches both edges. Finally, slant was the only parameter with an inverse sigmoid shape, where the slope is more horizontal around the middle and curls down and up on the lower and higher ends of the range, respectively.

[6]The values of the Bayes Factors for each model and parameter are in Section D.6.1

7.3 Q3: Distinguishable Levels

Because our most sophisticated model (Formula 6) fits average response as well as the magnitude of error depending on the level of the stimuli, we can use this model to calculate the levels of stimuli that will minimize the chance of one stimulus being confounded with another (see Appendix D.5). The bottom tetragrams of Figure 6 display the estimated optimal levels of stimuli (diamond marks) for 7, 5, 3 and 2 classes (from top to bottom), in each of the parameters. Columns labeled ϵ_7, ϵ_5, ϵ_3 and ϵ_2 in Table 2 report the probability (out of 1) of mistaking a stimulus for another for these configurations of 7, 5, 3 and 2 classes respectively. These calculations assume that each class is equally likely. The distribution of levels generally seeks to concentrate more stimuli wherever the curve is more vertical and, simultaneously, there is less variance. Note that this measure of confusion is based on a task in which a viewer looks at a word rendered in one level and has to make an absolute judgement of this stimulus using only the extremes of the scale as reference. For tasks in which viewers compare two levels of stimuli—much easier—see Experiment 2 below. The ranks of error follow similar patterns to the noise measurements of Subsection 7.1 above, especially for the largest numbers of classes: weight is best, followed closely by slant, then width, xHeight and then with contrast, serifs and aperture further behind.

7.4 Q4: Participant Population Variability

Because the model of Formula 6 parameterizes the participant's individual performance as a multiplier coefficient of the standard deviations (noise) at each stimulus level, we can characterize the expected variation in perceptual noise across our participant pool (Column $PC\sigma$ in Table 2). In other words, $PC\sigma$ models the spread in performance among participants, with some participant p's perceptions less noisy than the average participant—$pMult[p] < 1$–and other participants showing more noise instead—$pMult[p] > 1$.

These estimations are fairly uncertain due to the higher hierarchy of this parameter, but we observed clear differences between an unstable group of parameters (aperture, serifs and slant) and a stable group (contrast, weight, width). The 95% High Density Intervals barely overlap across groups or not at all (95% Column in Table 2). xHeight is in the middle.

7.5 Experiment 1: Discussion

Overall, slant and weight are the parameters that offer least perceptual noise for information encoding, followed by width and xHeight. Contrast, serifs and aperture are at the bottom of our list (Q1).

The shape of the perceptual response curves for weight and width (Q2) are expected and roughly consistent with Weber-Fechner's and Steven's laws [Fechner 1948; MacKay 1963] from Psychophysics: the larger the stimuli, the larger the difference has to be between levels of the stimuli to yield different responses. This is also consistent with existing practices in type design, where designers select glyph thickness for different weight variants of the same font family based on exponential or logarithmic curves so that they appear to be at *optically uniform* intervals [de Groot 2020]. This is, essentially, a natural way to compensate for the shapes that we observed, although based on the trained eye of type designers.

Table 2. Experiment 1 results. **R** refers to calculations based on Raw data, **M1** to the simpler model of Eq. 1 and **M2** to the model of Eq. 6, which models participant noise independently and different noise for each measured level of stimuli. **A** refers to calculations using the by-participant-**a**veraged data, **U** refers to **U**naveraged data (see Section 6.4.1). $PC\sigma$ is standard deviation of the per-person standard deviation multiplier coefficients ($pMult$—see 6.4.1 and 7.4). $\epsilon_{2,3,5,7}$ are the expected errors with optimally chosen inputs values for 2,3,5 and 7 classes respectively. Each column labeled "Rk" refers to the typographic parameter ranking of the parameter of the column to its left. The column labelled 95% reports the 95% High-Density Interval of the column to its left.

Typ. par.	$\sigma_{R,A}$	Rk	$\sigma_{R,U}$	Rk	Model	M1,A pars.	$\sigma_{M1,A}$	Rk	M2,U pars.	$\sigma_{M2,U}(x)$	$\overline{\sigma_{M2,U}}$	Rk	$PC\sigma$	95%	Rk	ϵ_2	Rk	ϵ_3	Rk	ϵ_5	Rk	ϵ_7	Rk
weight	0.078	1	0.11	1	quad.	$\beta_0 = 0.048$ $\beta_1 = 0.93$ $\beta_2 = -0.19$	0.079	1	$\beta_0 = 0.057$ $\beta_1 = 0.97$ $\beta_2 = -0.25$	0.068, 0.11, 0.14, 0.13, 0.14, 0.12, 0.13, 0.12, 0.099	0.12	2	0.24	0.17 0.32	3	0.000	1	0.001	1	0.072	1	0.17	1
width	0.08	3	0.14	3	quad.	$\beta_0 = 0.081$ $\beta_1 = 0.81$ $\beta_2 = -0.3$	0.08	2	$\beta_0 = 0.088$ $\beta_1 = 0.86$ $\beta_2 = -0.31$	0.12, 0.14, 0.15, 0.14, 0.14, 0.15, 0.14, 0.14, 0.16	0.14	3	0.24	0.17 0.32	2	0.000	2	0.018	2	0.15	3	0.24	3
contrast	0.11	5	0.19	5	cubic	$x_{c0} = -0.35$ $y_{c0} = -0.24$ $a = 0.41$ $c = 0.4$	0.12	5	$x_{c0} = -0.49$ $y_{c0} = -0.32$ $a = 0.3$ $c = 0.37$	0.17, 0.17, 0.18, 0.21, 0.22, 0.21, 0.21, 0.19, 0.17	0.19	5	0.22	0.16 0.29	1	0.003	5	0.12	5	0.28	5	0.35	5
xHeight	0.088	4	0.15	4	sigm.	$x_{s0} = -0.022$ $\kappa_s = 7.5$ $r_s = 0.57$	0.09	4	$x_{s0} = -0.019$ $\kappa_s = 8.3$ $r_s = 0.56$	0.14, 0.14, 0.14, 0.15, 0.15, 0.16, 0.15, 0.16, 0.16	0.15	4	0.3	0.21 0.4	4	0.000	3	0.029	4	0.17	4	0.26	4
slant	0.079	2	0.12	2	invsig	$y_{i0} = -0.013$ $\kappa_i = 3.7$ $r_i = 0.51$	0.083	3	$y_{i0} = -0.005$ $\kappa_i = 3.4$ $r_i = 0.51$	0.14, 0.13, 0.11, 0.067, 0.032, 0.075, 0.11, 0.15, 0.18	0.11	1	0.6	0.4 0.92	7	0.002	4	0.023	3	0.12	2	0.2	2
serifs	0.13	7	0.21	6	sigm.	$x_{s0} = -0.004$ $\kappa_s = 6.5$ $r_s = 0.55$	0.13	7	$x_{s0} = -0.011$ $\kappa_s = 9.9$ $r_s = 0.35$	0.24, 0.23, 0.23, 0.24, 0.21, 0.20, 0.22, 0.20, 0.21	0.22	6	0.49	0.31 0.77	6	0.015	6	0.13	6	0.29	6	0.35	6
aperture	0.13	6	0.25	7	sigm.	$x_{s0} = 0.009$ $\kappa_s = 4.4$ $r_s = 0.35$	0.12	6	$x_{s0} = 1e-04$ $\kappa_s = 7.1$ $r_s = 0.21$	0.24, 0.26, 0.25, 0.23, 0.25, 0.24, 0.25, 0.23, 0.25	0.24	7	0.42	0.28 0.62	5	0.16	7	0.31	7	0.4	7	0.43	7

A plausible explanation for the shape of contrast is that it behaves as the weight parameter, but with glyphs only partially reducing their thickness (mostly only vertical strokes). The inverse shape is due to how contrast was mapped to the stimuli scale: increasing contrast results in vertical strokes of diminishing weight.

xHeight shows a sigmoid shape that flattens at the beginning and the end of the range. This behavior took us by surprise because one would expect the small variations at the beginning and end of the range to be more distinct because differences in the height of the letter's body and the ascenders' protrusion above low letters are proportionally larger at both ends (Weber-Fechner's law). Instead, the slope at the middle of the curve is more pronounced, indicating better ability to appraise stimuli there. We speculate that this could be due to perceptual tuning to the more familiar middle range of this parameter (fonts with very high or low xHeight are unusual), or to floor and ceiling effects due to the finite coverage of range and how participants only saw the extreme ends of the "legend" in the stimuli presentation. The curves for serifs and aperture are similar; however, the increased noise in these two parameters and the smaller prominence of the flattening—the aperture curve is almost completely straight—makes these effects less practically relevant.

Finally, the inverse-sigmoid shape of slant is consistent with psychophysical measures of orientation perception; judgment bias and error increase as orientation deviates from straight (0°, 90°, 135°and 180°) angles [Durgin and Li 2011; Howe and Purves 2005].[7]

This characterization of the curves, which is different for each typographical parameter, will allow designers of visualizations to use non-linear mappings that compensate for the systematic bias in perception, beyond how type designers currently compensate weight in their fonts.

If used for categorical ordinal representations (Q3), we predict that aperture might be just acceptable to represent a binary variable (16% expected error), with all other typographic parameters showing very few expected errors in this case (< 2%). For 7 levels, weight and slant offer levels of error that might be acceptable in many applications (both < 20%). Note that optimal levels and accuracy will change if the mapped data variable does not have a uniform frequency distribution.[8] We also learned that people's perception error was less stable across our population for slant, serifs and aperture (Q4).

8 EXPERIMENT 2: JUST NOTICEABLE DIFFERENCES

Unlike Experiment 1, this experiment measures the observer's ability to differentiate levels of a typographic parameter when words rendered with these different levels are simultaneously visible. That is, we measure the Just Noticeable Differences (JNDs) along each typographic parameter.

8.1 Task

Participants saw two renderings of different words on the screen, each with a different level of the tested parameter. All other parameters were kept at the neutral level. A prompt asked the participant to select the word with the higher level of the parameter with a radio button (see Figure 9). For example, the prompt indicated "select which word has a higher weight". The correct answer was the left or the right in an unpredictable order. This follows a *two-alternative*

[7]Note that the range of slant tested is only a small local interval around 90°.

[8]Optimal levels for other circumstances can still be easily calculated with the code from the supplementary materials.

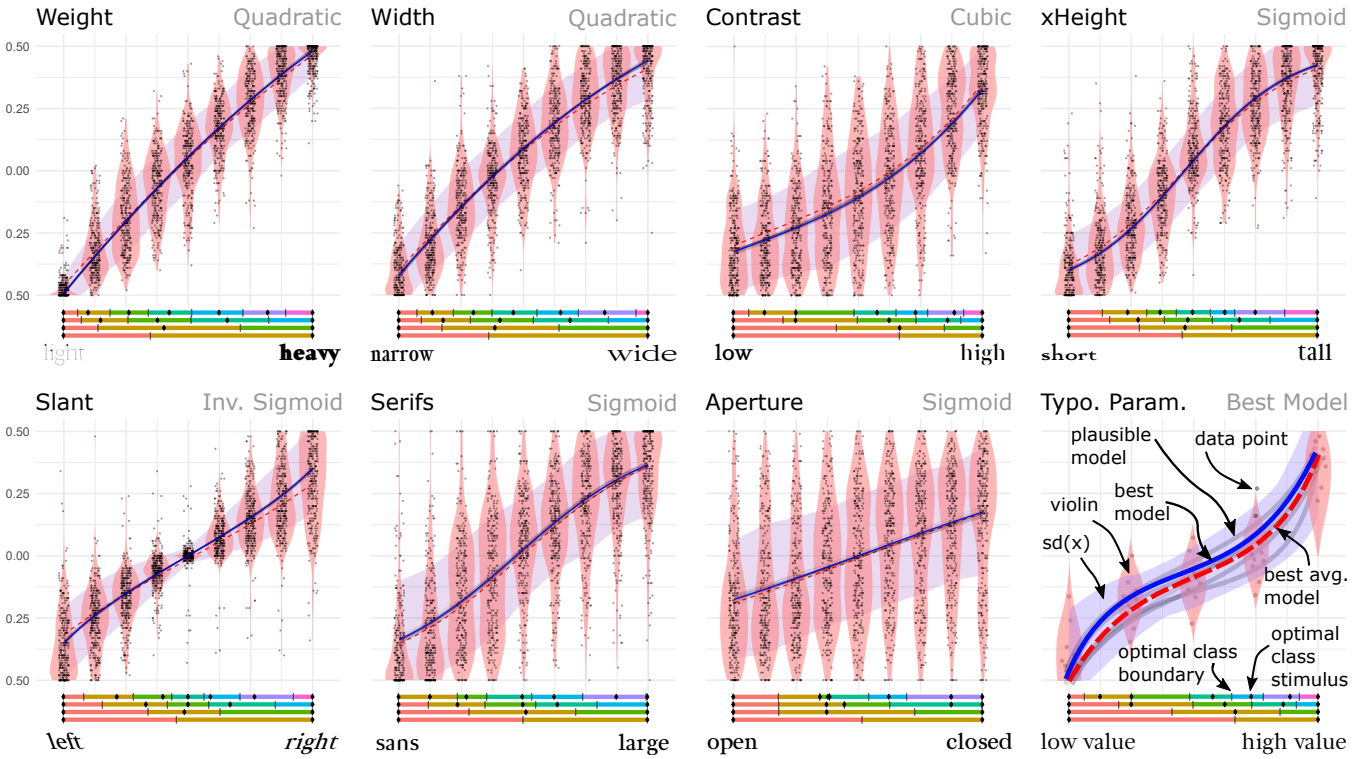

Fig. 6. Experiment 1 unaveraged models. The Bottom right figure is the legend for the others. Width of violin plots indicates density of data points (normalized within input value). Figure represents individual trial data points by all participants for that parameter. Plausible models from the simulations drawn = 30 (grey transparent lines). These are drawn randomly from MCMC simulation and suggest certainty of the model chosen. Bottom panel displays optimal stimulus points and boundaries for seven, five, three and two classes from top to bottom.

forced choice task (2AFC) paradigm. Text at the bottom of the page indicated the trial's number and the total number of trials.

8.2 Procedure

Participants provided informed consent. Then they saw instructions on how to perform the task and a detailed explanation of the corresponding typographic parameter. We measured JNDs at five equidistant levels covering the parameters' range. For each level, the experiment presented staircase procedures that approached the level from above or from below the scale, except for the lowest end of the range, which was only approached from above, and the highest end, which was only approached from below. There were thus eight different staircases. Testing on each of the staircases was interleaved in an unpredictable fashion to reduce carryover effect from one trial to the next. Participants made 35 judgements on each staircase, for a total of 280 trials. The efficient staircase was adaptive and based on Kontsevich et al.'s Bayesian procedure which calculates the next level to test by maximizing the potential information gain [Kontsevich and Tyler 1999]. Each of the staircases results in a measure of the Just Noticeable Difference (JND) at that level (from above or below), which estimates the stimulus level at which the psychometric function crosses half way between chance level of error and perfect performance (i.e., the stimulus at which the likelihood of error is 25% in a 2AFC task—more detail in Appendix E.2). Due to differences in the implementation of the different typographic parameters, the granularity of discretization of the range is different for each (these are displayed in the rightmost column in Table 1).

During pilot testing we realized that many participants were confused by a task asking them to choose the rightmost slant if both slants were left sided or when the positions of the slanted words were reversed. To maintain consistency of the prompt and avoid noise due to misinterpretation, we separated left slant and right slant treating them as two different parameters.[9]

8.3 Participants

We obtained 285 full sets of trials for this experiment, of which we kept 237 (38 were not complete and 48 were filtered). Participants did not repeat the same parameter but could do multiple parameters (105 unique participants completed the 237 full sets included in the analysis). The only parameter for which we did not reach 30 full sets was aperture, for which we tested 53 participants but only 20 complied with the acceptability conditions defined a priori. 40 participants were female, 64 male, and one did not disclose sex.

[9]This provides measurements at additional levels for this parameter (16 staircases instead of 8). We report separate or joint results when appropriate.

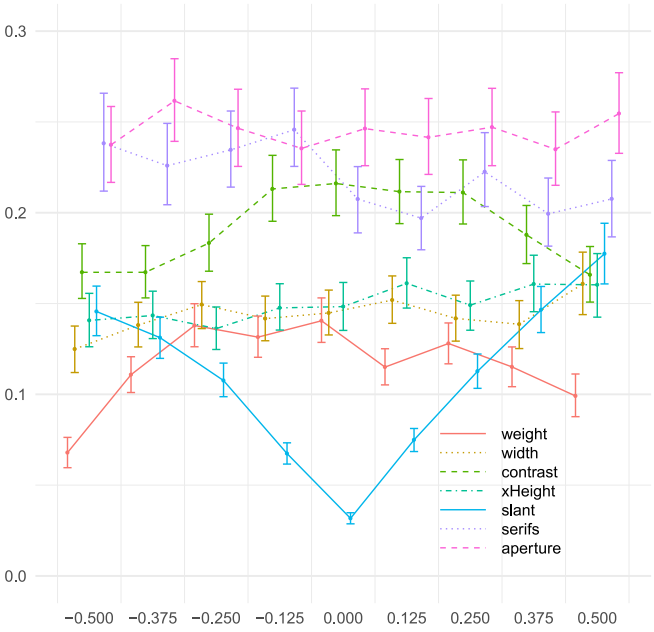

Fig. 7. The standard deviations (y axis) for each input level of all parameters. Error bars indicate 95% High Density Intervals, which is a measure of the uncertainty of the estimation. Horizontal position is slightly shifted to avoid visual overlap of the parameters.

Fig. 8. Average standard deviation across all 9 parameter stimuli points, for each typographic parameter. Error lines indicate 95% HDI.

Fig. 9. Experiment 2 task (weight parameter).

Participants ranged between 18 and 61 in age, with a median of 31 and a mode of 25. There is further detail in Appendix E.1.

8.4 Specifics of the Analysis

The analysis of this experiment incorporates elements from the analysis in Experiment 1 as well as from Kay and Heer's approach in their Bayesian analysis of JNDs [2016]. Unlike in Experiment 1, we do not compare different curves for plausibility, and we simply fit a cubic curve with four parameters that should provide a reasonable account of variation as it depends on the stimulus level. This analysis decision was made partially because the data from this experiment covers fewer levels of the stimuli range and because there are fewer repetitions (each staircase provides a single estimation of JND, despite consisting of 35 trials).

Like Kay and Heer [2016], we fit a log-linear transformation of the cubic curve. This is consistent with the observation that, in our data (as in theirs), the size of residuals grows with the measures of JND (i.e., the larger the magnitude of the measured JND, the larger the variability). We account for the approach measure of JND from above or below as an additional term in the function that models the mean. This term fits the separation between the two ways to approach the stimulus (see Equation 8); this allows us to fit a single model for both approaches, but avoids having to average them or to apply a procedure such as Harrison et al.'s [2014].[10] The noise model is expressed by the log-normal probability in Equation 7, where the mean (Equation 8) depends on the stimulus level (x), the approach (from above or below, ap) and the participant (p).

$$P(y \mid x, ap, p) = \exp\left(\frac{1}{\sigma(x)\sqrt{2\pi}} e^{-(y-\mu(x,ap,p))^2 \Big/ 2\sigma(x)^2}\right) \tag{7}$$

$$\mu(x, ap, p) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \delta ap + \gamma p \tag{8}$$

In Equation 8 the additive δ fitted coefficient expresses the differences between approaches from above and below explained above. The γ coefficient captures the variation between individual participants. The variance ($\sigma(x)$) is modeled independently for each stimulus level as in Equation 6. Unlike in Experiment 1, here we cannot model variation within the individual participant's response because we only gather a single JND value per unique combination of participant, stimulus level and approach.

As in Experiment 1, we apply censoring to estimate variance and averages beyond the measurable range. This is particularly important here because the staircases have different measurement ranges dependent on stimulus level and approach (see Appendix E.4).

9 EXPERIMENT 2: RESULTS

We report the results according to the four questions in Section 4. Table 3 summarizes the main outputs of the analysis.

9.1 Q1: Perceptual Noise

Just Noticeable Differences are measurements of perceptual noise because they estimate the magnitude of change in the stimulus that is necessary for the viewer to reliably detect the change (i.e., JND is an implicit measure of signal-to-noise ratio). Averaged across all stimulus levels per typographic parameter (Column $\overline{JND_R}$ in Table 3), the smallest JNDs correspond to weight, followed by slant at some distance, with width and xHeight after, then serifs and contrast, and with aperture clearly behind. The average JND magnitudes calculated from the fitted models are similar; only width and xHeight swap places in the ranking.

Notice that the measure of slant is aggregated from two independent sets of measures done on different sets of participants.

[10] We considered using Harrison et al.'s procedure, in which the curve fits JND measurements shifted in stimulus (x axis) by the average JND. Although this makes conceptual sense, it simply does not fit our data well in qualitative posterior predictive checks.

Table 3. Main result table for Experiment 2. **R** refers to calculations performed on the raw data, whereas **M** refers to the model-based calculations. σ_{part} measures the variability of γ in Equation 8. The column to its right is the 95% HDI of this parameter. Each column labeled "Rk" refers to the typographic parameter ranking of the parameter of the column to its left.

Typ. par.	$\overline{JND_R}$	Rk	$\overline{JND_M}$	Rk	Model pars.	σ_{part}	95% range σ_{part}	Rk	Num JND Intervals	Rk
weight	0.04	1	0.033	1	$\beta_0 = -3.4\ \beta_1 = 0.92\ \beta_2 = -0.28\ \beta_3 = 0.28$	0.45	0.33 — 0.6	3	32	1
width	0.1	3	0.088	4	$\beta_0 = -2.4\ \beta_1 = 0.74\ \beta_2 = -0.34\ \beta_3 = 0.83$	0.37	0.27 — 0.49	1	11	4
contrast	0.19	6	0.16	6	$\beta_0 = -1.9\ \beta_1 = -1.1\ \beta_2 = 0.71\ \beta_3 = 1.9$	0.42	0.3 — 0.56	2	7	6
xHeight	0.11	4	0.085	3	$\beta_0 = -2.6\ \beta_1 = 0.15\ \beta_2 = 1.2\ \beta_3 = 2.2$	0.54	0.4 — 0.71	6	11	4
slantL	-	-	-	-	$\beta_0 = -2.2\ \beta_1 = -1.8\ \beta_2 = 2.2\ \beta_3 = 6.6$	0.48	0.35 — 0.63	4	-	-
slantR	-	-	-	-	$\beta_0 = -2.0\ \beta_1 = 1.4\ \beta_2 = 0.84\ \beta_3 = -4.2$	0.5	0.38 — 0.65	5	-	-
slantCombined	0.086	2	0.074	2	-	-	-	-	15	2
serifs	0.18	5	0.14	5	$\beta_0 = -2.0\ \beta_1 = 0.88\ \beta_2 = 0.6\ \beta_3 = 1.0$	0.61	0.46 — 0.8	8	8	5
aperture	0.32	7	0.25	7	$\beta_0 = -1.4\ \beta_1 = 0.32\ \beta_2 = -0.19\ \beta_3 = -1.7$	0.58	0.41 — 0.79	7	4	7

Combining the two slant sub-ranges expands the range measured and hence halves the JNDs of the individual subtests.[11]

Figure 10 is analogous to Experiment 1's Figure 8 and shows the stability of the averaged JNDs for the different parameters (HDI intervals calculated from the MCMC chains), except for the combined slant parameter, which only shows the average.[12] The only intervals that overlap are width and xHeight, which supports the rankings at the beginning of this subsection.

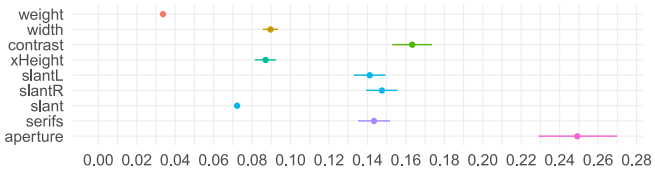

Fig. 10. Average JND estimations across all levels for each parameter. Error lines represent 95% HDIs.

9.2 Q2: Curve Shapes

Figure 11 shows the best fitted curves of the magnitude of JNDs per parameter. Figure 12 presents the curves of all the typographical parameters for better comparison, with the left and right slant curves amalgamated into one. Visual inspection of the figures and the best parameter values fitted provide a strong indication that weight and width are roughly linear with a positive slope. Serifs shows increasing variability as the serif size increases, with a larger increase towards the end and xHeight has stable (and fairly small) JNDs up to the end of the scale. Slant becomes less differentiable as the angle deviates from the vertical, but there is a moderate increase of noise just around the upright level. The pattern is quite similar for both left and right slant. Contrast and aperture are generally quite noisy.

[11] Individual JNDs are defined according to the variation range of the subrange; when integrated into the double range, a JND with respect to the overall range is halved.
[12] The aggregated HDI of slant cannot be calculated from two independent MCMC chains, but the means is reasonably approximated as the mean of the separate slantL and slantR measurements, which have the same number of samples.

9.3 Q3: Distinguishable Levels

The bars at the bottom of each subfigure in Figure 11 indicate the distributions of stimuli levels that contain the largest number of points that are separated by at least a JND (see Appendix E.5 for the algorithm and rationale). This guarantees a conservative upper bound of 25% confusion error with neighboring levels of each of the categories. The last two columns in Table 3 summarize the results for each parameter: weight is best, with more than double the number of categories (32) than the combined slant (15). Width and xHeight are tied at 11, followed by serifs, contrast, and aperture, where only four categories would be distinguishable with this level of error.

9.4 Q4: Participant Population Variability

The estimated range of variability of the population in this experiment does not offer a clear differentiation between parameters. Ranges of σ_{part} (column 8 in Table 3) are wide and clustered around the same standard deviation range (between 0.37 for width and 0.61 for serifs). The substantial overlap indicates insufficient reliability of the estimation. This is likely due to the reduced number of data points (one per participant and level), which increases uncertainty in the estimation of variability.

9.5 Experiment 2: Discussion

The JND measurements (Q1) place weight by far as the most differentiable typographic parameter, followed by slant (when left and right slants are combined), then xHeight and width very close to each other (and statistically indistinguishable), and serifs, contrast, and aperture at some distance. The JNDs of aperture are much larger than for any of the other parameters, and almost an order of magnitude larger than those of weight. Moreover, the measurements on aperture are likely to be over-optimistic due to the rejection of a large number of participants whose generated data did not comply with our pre-set conditions despite them having tried their best. The aperture parameter, despite combining typographic variations in both aperture and junction, seems too subtle for most people and filtering criteria might have been too strict.

The experimental design in Experiment 2 measures fewer levels than in Experiment 1, which means that we can make fewer distinctions regarding the shapes of the perceptual noise curves (Q2), yet we can still make some observations as to the general increasing or

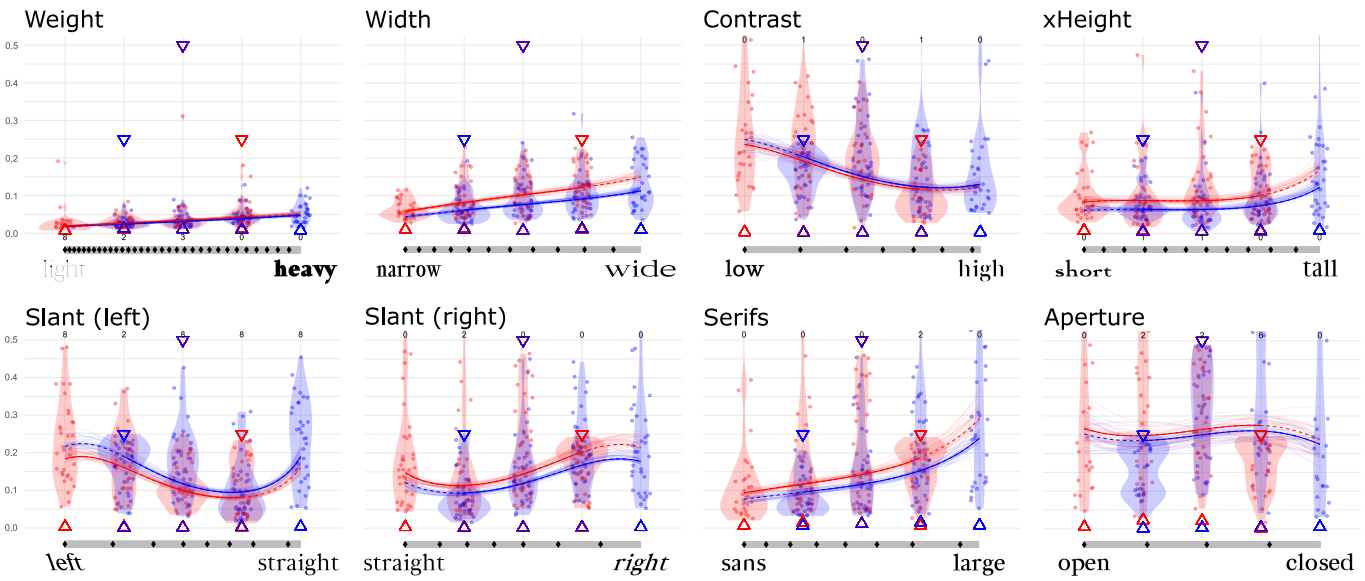

Fig. 11. Fitted curves of Experiment 2. Red and blue elements correspond to approaches from above and below respectively. Solid lines show the most likely curves fitted. Discontinuous curve segments indicate that the curve is extrapolated (there is no data on approaches from above at the rightmost level of the range). Transparent curves show random samples of models from the MCMC, which indicate the uncertainty of the fit. Downward- and upward-pointing triangles show the measurement bounds of the staircase at that level, which were used for censoring (see more detail in Appendix E.4). The violin display the distribution of the individual samples at each level, which are plotted as dots in the corresponding color, with a small jitter in the horizontal position for visibility. The bar with diamonds at the bottom shows stimuli levels that are separated by at least a JND.

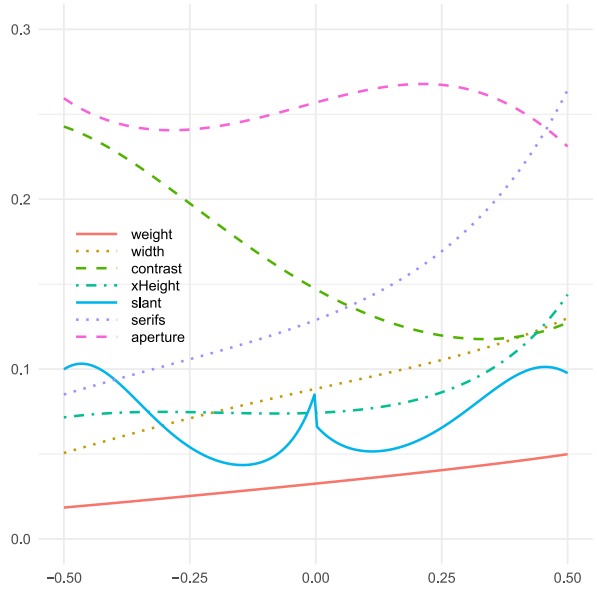

Fig. 12. Fitted curves of JND values for Experiment 2. Only central estimated values are shown, for clarity. The two slant experiments are combined into a single curve, hence the small discontinuity in the middle. Combined slant might seem low in comparison with the individual curves in Figure 11 because the vertical axis values are halved when the two datasets are combined to each cover half of the horizontal range.

decreasing pattern. Weight and width's shapes (increasing value of JNDs from left to right in weight and width) are consistent with the curve shapes from Figure 6 in Experiment 1 (i.e., the just noticeable differences increase as the perception curve becomes flatter), and therefore also roughly consistent with Weber-Fechner's law, despite the fact that the tasks are different. As in Experiment 1, contrast has an inverted pattern, although much noisier than weight and width. xHeight's curve is mostly flat except towards the high end of the range—i.e., consistent with Experiment 1's result on the high end, but not on the low end. Serifs shows a monotonic increase in JNDs that is only weakly consistent with the results from Experiment 1. The shape of the combined slant, for which we have double the levels of samples (see Section 8.2), shows a large dip in the middle (angles close to upright text) consistent with the noise measurements of Experiment 1, but with an unexpected twist: the JND close to the upright position is larger than the next immediate one. For aperture, which has fairly high JNDs, we cannot make strong claims about the shape of the curve because the measurements are too close to the upper bounds (the downward pointing triangles in Figure 11), which makes the measurements more uncertain.

Dividing the range space into JND intervals (Q3) gives us a surprisingly large number of discrete levels that could be mapped to many values of a continuous or discrete variable. Even serifs—one of the parameters with the largest JNDs—can offer up to 8 different levels separated by at least one JND of distance from each other. This is encouraging for encoding information, but it is important to remember that JNDs, by definition, still would incur an average of 25% error. Additionally, the JND task is representative of almost

optimum conditions for parameter judgement: words are rendered next to each other, and the viewer can quickly direct their gaze back and forth between the (different) words to find differences. In practical terms, this suggests that aperture can only reliably represent a binary class distinction, whereas we can be confident that even contrast or serifs can represent ordinal data with at least four categories (probably more). Weight, slant, width, and xHeight have JNDs small enough to reasonably encode continuous variables.

Our experiments measured slant separately for right and left leaning renderings. The main rationale was to acquire JND measurements that avoid noise introduced by the experimental paradigm itself in the presentation of the task. Although this gives us more internally valid JND measurements, it does not account for the fact that people might actually experience similar confusion when encountering slant encodings in text or in visualizations. This might be further compounded if the text in visualization is rotated. Although further empirical confirmation is required, this should serve as an early warning that applying slant can be problematic because the inherent spatial-orientation aspect of the parameter variation could interfere with other spatial aspects of the representation (e.g., it might place additional requirements on mental rotation). If only one side of slant is available (i.e., range is halved), then this parameter only offers more differentiability than contrast and aperture.

Interestingly, the left slant and right slant measurements were very similar (when mirrored). This is somewhat surprising, since right slant is far more common in text than left slant (italics most often have right slants). Thus, familiarity might not play too big of a role in perceptual differentiability, at least for this parameter. Also interesting is that, according to our measurements, the most sensitive regions of the slant parameter (the smallest JNDs) are not immediately around the vertical slant, but offset by about 5° in each direction. This goes against evidence from the literature on perception of orientation, which suggests that the human perceptual system is most sensitive to departures from horizontality and verticality [Durgin and Li 2011]. We can only speculate that this might be caused by the additional visual complexity and variation of orientation contained in lettershapes as compared to the isolated simple segments from traditional experiments in Psychophysics.

10 GLOBAL DISCUSSION

In this Section we interpret the combined results from both experiments, examine the applicability of the results in real scenarios and discuss the limitations and future avenues for this research.

10.1 Results Summary

Although the tasks of Experiment 1 and 2 are distinct, the parameter rankings from both experiments are similar. This suggests at least a partial common latent source of noise associated to the perceptual characteristics of each parameter. The rankings classify parameters in roughly four groups: weight and slant are least noisy, followed by xHeight and width, then contrast and serif, and with aperture-junction by itself as the noisiest parameter (Q1). Within each of these four groups the parameters are close or swap places.

Readers should interpret the more accurate perception in the JND experiment results as a lower bound measure of perceptual noise that corresponds to close-to-ideal perceptual conditions. We expect JNDs to match perceptual tasks that involve comparison of text in close proximity. For example, JNDs should approximate well whether people notice increases of a parameter in subsequent words of a sentence. The absolute perception task from Experiment 1 is harder, shows generally larger error (closer to an "upper bound" for each parameter), and corresponds to tasks that require reading a value from its encoded typographic parameter, possibly with a legend. Other tasks such as finding extrema values, or comparing values across locations will likely fall somewhere in between.

The modeling of noise along the range of each parameter provides conclusive evidence that their perception is non-linear (Q2). The per-parameter response curves are not very surprising for weight and width (they show the flattening towards the right predicted by Weber-Fechner's and Steven's laws). The other parameters, which have not been tested before, show are a wider variety of distributions. Most would be difficult to predict only from existing results in Psychophysics. The measurements of level of noise from both experiments, combined with the variation of the curve's slope, provide the most complete characterization of these perceptual spaces to date. In some cases (e.g., the edges of the range in slant) slope and error counteract each other. In contrast, sometimes the regions with steep slopes have less variability (e.g., the *light* and *narrow* ranges of weight and width respectively).

The analysis also gives an estimation of how much perceptual variability to expect from a heterogeneous population. This should be taken only as an initial approximation because these are second-order estimations with high uncertainty. Nevertheless, the statistical analysis of Experiment 1 provides sufficient data to indicate that weight, width and contrast perception were more homogeneously perceived across our participant sample (Q4).

We interpret our results as supportive of Brath and Banissi's vision of varied and widespread infotypographic applications [2014a]. Several of the parameters offer substantial ranges of discriminability for categorical and continuous mapping of information attributes.

10.2 Applicability of the Results

Probably the most immediate application of this work is in the selection of the appropriate typographic parameters to represent data as part of text within visualizations and infographics. If accuracy matters, weight and slant will be most differentiable and appropriate to map to continuous variables, whereas serifs and aperture are subtle enough that they probably should not be used for more than two or three distinct categories. Regardless of whether the data variables are categorical/ordered or continuous, our characterization of the perceptual responses (Q2) supports linearizing the range of each of the parameters through the corresponding inverse functions. If the variable is categorical, this will help select the levels that minimize confusion (Q3). We provide these selections for sets of two, three, five and seven levels (based on Experiment 1 data), and for distribution of non-overlapping JNDs (based on Experiment 2 data).

A related application is the typographic design of document templates. Our results can help select formatting conventions for headings and highlighting that make the hierarchies more distinguishable. Similarly, designers of static fonts can use our results

to select perceptually-normalized levels in seven parameters. For variable fonts, pre-normalizing the parameters with the empirically-generated curves will presumably help make the space of possible font instances more perceptually uniform.

There are some important considerations for those looking to apply our results. A key one is the granularity of infotypographic mappings. Our results reflect accuracy and judgements at the level of words. All letters in a word had the same levels. This is appropriate in many applications because words are the most common semantic unit. However, some applications might require variation of parameters at the morpheme, syllable or even letter level. In these cases we anticipate noise because some single letters do not have sufficient features to assess a parameter (e.g., "l" is very simple).

Some parameters also have specific applicability limitations. For example, xHeight will be harder to decode through a word without letters with ascenders or descenders (i.e., if it is spelled only with aceimnorsuvwxz). The use of the slant parameter that we tested also assumes that the word is placed horizontally. Because perception of angle is dependent on the viewer's frame, using angle in a visualization with rotated words might result in irregular or unpredictable perceptual patterns. In exchange, slant is the only parameter we tested that has a practical and natural neutral point (the upright position). Slant will be useful to represent data that has ranges spanning from negative to positive and where zero point is significant.

Applications requiring that a word's footprint is constant regardless of the level of the typographic parameter should not map to width, and weight would require additional work such as special inter-letter spacing. Some parameters also vary the density of ink or black pixels more than others. This will affect the salience of words. Variations of weight, but also width and xHeight, are more conspicuous or visible at the overview level than those of slant, contrast, serifs and aperture. Note that salience and discriminability are not necessarily linearly or monotonically related, and there are applications in which it might be desirable to represent data within the text in a way that is unobtrusive and inconspicuous until the viewer decides to explicitly access the encoded information. For example, we might want to encode data in text that is meant mostly for normal reading and only occasionally requires reading of the encoded data. Contrast or serifs would work for this. Nevertheless, it should be clear that our experiments do not measure salience; this topic deserves further study (see also Strobelt et al.'s work [2016]).

We deliberately excluded color and size and focused on strictly typographic letter shape parameters to keep the studies manageable. However, many desirable applications of infotypographic mappings might also rely on the use of color and size, which are extremely common visual channels (Alexander et al. comprehensively examine font size and its biases [2018]). Although color and size will undoubtedly be useful along InfoTypography, they may interfere with the perception of the parameters we tested. For example, a lighter color might make weight less distinguishable but not width.

Font alterations might also affect recognition times and readability, which we did not study [Sanocki 1987, 1988].

Finally, some of the parameters will have semantic associations to the data that they represent for particular applications. For example, to encode speech volume in a libretto for musical theater, weight seems straightforward. Others have suggested the use of left slant to represent irony or sarcasm [Houston 2013]. The discovery and design of these relationships is an open research avenue.

10.3 Limitations, Threats to Validity and Future Work

Any experimental design imposes interpretation restrictions. A major limitation of our study is that it tested each parameter individually (when we varied one parameter, all other parameters were kept constant at their neutral values). Although this still allows us to compare each parameter's perception, it offers only a crude approximation for multi-variate scenarios. Some typographical parameters will likely not be separable; naïvely extrapolating our data will likely miss non-additive effects. For example, contrast is almost imperceptible in the low end of the weight range. More research is required to map these second-order interactions and we do not recommend the use of our data to estimate multi-parameter differentiability.

Our tests were also carried out with a specific variable font. The design of variable fonts is diverse, and the more different from our model, the more likely our results will be inaccurate for a font. Nevertheless, we believe that our John Fell modification is representative of commonly used fonts that support reading (e.g., Times New Roman or Calibri). We also provide the font samples, measures, and comparisons with existing fonts that allow font designers to extrapolate our results to help with their own designs. Two related generalization issues are that we tested only lowercase letters (we are interested in supporting reading scenarios), and that the Latin alphabet is, naturally, only one of many used worldwide.

We carefully selected the typographic parameters for our experiment. Among the variable parameters featured in v-fonts [Sherman 2021], weight, width and slant (sometimes labeled as *italics*) are dominant. Nevertheless, there are infinite ways to organize morphological variations of letter shapes, and some other parameters (familiar or unfamiliar) might be more appropriate for infotypographic purposes than ours. For example, a chunkier slab serif parameter might offer better accuracy than ours.

In our experiments we did not have complete control of the size, resolution or visual angle of the displayed words of the experiment; people could view the screen at whichever distance they preferred. We chose to display the stimuli at a size likely to enable reasonable estimations of the parameters. Displaying words in a larger size or resolution might increase the reader's performance—there exists some evidence that larger text results in shorter eye fixations [Rello et al. 2016] and that size affects reading speed [Legge and Bigelow 2011], but reading speed might not correlate with value decoding. We suspect that perception will become much noisier as the text becomes much smaller, especially when the features of letter shapes become harder to see due to the resolution of the display or the acuity thresholds of the human fovea. Fortunately, digital display resolution keeps increasing and the human acuity threshold is well below typical reading font sizes (8 to 12pt) at regular reading distances (up to 2'—60cm [Brown 2020]). Nevertheless the effects of font size is still an active area of research (e.g., [Chaparro et al. 2010; Dobres et al. 2016; Rello et al. 2016]), and further research is needed to determine the acceptable size thresholds for infotypographic use.

Another factor that is certain to influence viewers' accuracy in decoding infotypographic parameters is time spent examining the

letter shapes. We asked participants to achieve both accuracy and speed which, in practice, means that participants find their own speed-accuracy balance. Although it would be interesting to know to what extent accuracy can improve by encouraging more effort (e.g., by rewarding accuracy), we think that the 4.5 and 3.4 second median trial durations are good initial representatives of what would be acceptable in real-world applications.

Readers interested in the more subtle parameter of aperture should interpret the JND measurements carefully because these were obtained from a reduced number of participants, which is likely to overestimate its accuracy. This is due to how we discarded participants using *a priori* criteria. This parameter's high noise caused many participants to fail the acceptability criteria because they were indistinguishable from participants randomly clicking. Hence, the participants in that parameter's sample represent the most capable of judging it. This combination of aperture and junction might just be below the bar of utility to represent binary class data, unless people were extensively trained or the purpose of the encoding was precisely to hide data (as in Xiao et al.'s work [2018]).

Another limit of the JND experiment is that we were not able to accurately measure the slope of the Psychometric Function (see [Klein 2001]). This does not invalidate the results; the JND measures are still accurate, support parameter comparisons and will help chose intervals for the stimuli. However, it does prevent easy generalization of the results to an arbitrary number of ordered or categorical classes or to set an acceptable error threshold and work backwards from there. Both of these are feasible for Experiment 1.

Finally, there are important and exciting open exploration avenues in InfoTypography and letter shape perception. An obvious first direction is building a systematic map of the human perception of the multi-dimensional space of letter shape variation, analogous to work in color perception (e.g., CIE-LAB [Ohno 2000]). Although the prospect is daunting due to the higher-dimensionality of the space, it can open the door to a better design of documents and visualizations or lead to automated selection and design of fonts for specific applications. An additional question is whether humans are able to simultaneously integrate infotypographic encodings with the phonological and semantic information from the text. If integration of both types of information is natural or trainable, this can lead to new enriched modalities of reading.

11 CONCLUSION

InfoTypography is the use of type design parameters—such as the length of serifs and the contrast of letter shapes—to represent information. InfoTypography has recently become much more accessible due to the generalization and availability of variable fonts in web platforms, and offers great potential to complement existing representation techniques in infographics and information visualizations that contain text. This paper describes the results of two crowdsourced experiments designed to characterize human perception of seven infotypographic parameters: weight, width, contrast, xHeight, slant, serifs, and aperture/junction. The results show that different parameters have different levels of perceptual noise, with weight and slant having the least noise and contrast, serifs and aperture the most. The experiments also allow us to model the non-linear perceptual response to changes in the parameters which, in turn, will allow practitioners to optimize infotypographic encodings of continuous or ordinal variables. The results are applicable beyond visualization use and can support typographers with their choice of document fonts and designers of non-variable fonts with their choices of what font variants to create.

ACKNOWLEDGMENTS

The authors wish to thank Dr. Glenna Nightingale, Dr. John Thomson and Professor Steve Linton for mathematical advice, the VIXI Research Group at UVic, the St Andrews Systems team (fixit), and Jose Marques especially, for the timely technical help and patience, the members of SACHI at the University of St Andrews, Charles Perin, Teseo Schneider and Katherine Skipsey for reviews of the manuscript and advice, Timoth Sheridan for initial explorations, and the participants of the experiment for their effort. This work was supported by EPSRC (Infotypography grant EP/M008932/1 — UK) and by NSERC (Discovery Grant 2020-04401 — Canada). Thanks also to Prototypo.io for discussions and making their code available.

REFERENCES

Adobe Corporation. 1997. *Designing Multiple Master Fonts.* https://www.adobe.com/content/dam/acom/en/devnet/font/pdfs/5091.Design_MM_Fonts.pdf

Shehzad Afzal, Ross Maciejewski, Yun Jang, Niklas Elmqvist, and David S. Ebert. 2012. Spatial Text Visualization Using Automatic Typographic Maps. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (Dec. 2012), 2556–2564. https://doi.org/10.1109/TVCG.2012.264

Hirotugu Akaike. 1974. A New Look at the Statistical Model Identification. *IEEE Trans. Automat. Control* 19, 6 (Dec. 1974), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Eric Alexander, Chih-Ching Chang, Mariana Shimabukuro, Steven Franconeri, Christopher Collins, and Michael Gleicher. 2018. Perceptual Biases in Font Size as a Data Encoding. *IEEE Transactions on Visualization and Computer Graphics* 24, 8 (Aug. 2018), 2397–2410. https://doi.org/10.1109/TVCG.2017.2723397

Apple Corporation. 2020. *About Apple Advanced Typography Fonts.* https://developer.apple.com/fonts/TrueType-Reference-Manual/RM06/Chap6AATIntro.html

Aries Arditi and Jianna Cho. 2005. Serifs and Font Legibility. *Vision Research* 45, 23 (Nov. 2005), 2926–2933. https://doi.org/10.1016/j.visres.2005.06.013

Gantugs Atarsaikhan, Brian Kenji Iwana, Atsushi Narusawa, Keiji Yanai, and Seiichi Uchida. 2017. Neural Font Style Transfer. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Vol. 05. 51–56. https://doi.org/10.1109/ICDAR.2017.328

Samaneh Azadi, Matthew Fisher, Vladimir G. Kim, Zhaowen Wang, Eli Shechtman, and Trevor Darrell. 2018. Multi-Content GAN for Few-Shot Font Style Transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 7564–7573. https://doi.org/10.1109/CVPR.2018.00789

Elena Balashova, Amit H. Bermano, Vladimir G. Kim, Stephen DiVerdi, Aaron Hertzmann, and Thomas Funkhouser. 2019. Learning A Stroke-Based Representation for Fonts. *Computer Graphics Forum* 38, 1 (2019), 429–442. https://doi.org/10.1111/cgf.13540

Shumeet Baluja. 2017. Learning Typographic Style: From Discrimination to Synthesis. *Machine Vision and Applications* 28, 5 (Aug. 2017), 551–568. https://doi.org/10.1007/s00138-017-0842-6

Scott Bateman, Carl Gutwin, and Miguel Nacenta. 2008. Seeing things in the clouds: the effect of visual features on tag cloud selections. In *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia.* 193–202. https://doi.org/10.1145/1379092.1379130

Benjamin Bauermeister. 1988. *A Manual of Comparative Typography: The PANOSE System.* Van Nostrand Reinhold.

Sofie Beier. 2012. *Reading letters: designing for legibility.* Bis Publishers.

Sofie Beier, Katrine Sand, and Randi Starrfelt. 2017. Legibility Implications of Embellished Display Typefaces. *Visible Language* 51, 1 (April 2017), 112–132. https://adk.elsevierpure.com/da/publications/legibility-implications-of-embellished-display-typefaces

Jacques Bertin. 1967. *Sémiologie graphique: les diagrammes, les réseaux, les cartes.* Flammarion, Paris.

Richard Brath and Ebad Banissi. 2014a. The Design Space of Typeface (Poster). In *IEEE Conference on Information Visualisation VIS 2014.* Institute of Electrical and

Electronics Engineers (IEEE), Paris, France. https://openresearch.lsbu.ac.uk/item/8775v

Richard Brath and Ebad Banissi. 2014b. Using Font Attributes in Knowledge Maps and Information Retrieval. In *CEUR Workshop Proceedings*, Vol. 1311. CEUR Workshop Proceedings, London, 23–30. https://openresearch.lsbu.ac.uk/item/877qx

Richard Brath and Ebad Banissi. 2016. Using Typography to Expand the Design Space of Data Visualization. *She Ji: The Journal of Design, Economics, and Innovation* 2, 1 (March 2016), 59–87. https://doi.org/10.1016/j.sheji.2016.05.003

Richard Brath and Ebad Banissi. 2017. Multivariate Label-Based Thematic Maps. *International Journal of Cartography* 3, 1 (2017), 45–60. https://doi.org/10.1080/23729333.2017.1301346

Richard Brath and Ebad Banissi. 2019. Bertin's forgotten typographic variables and new typographic visualization. *Cartography and Geographic Information Science* 46, 2 (March 2019), 119–139. https://doi.org/10.1080/15230406.2018.1516572

Robert Bringhurst. 2004. *The elements of typographic style.* Hartley & Marks Vancouver.

Andy Brown. 2020. Viewing Distance and Font Size - Design Resources. http://resources.printhandbook.com/pages/viewing-distance-font-size.php

Barbara Brownie. 2007. One Form, Many Letters: Fluid and Transient Letterforms in Screen-Based Typographic Artefacts. *Networking Knowledge: Journal of the MeCCSA Postgraduate Network* 1, 2 (Dec. 2007). https://doi.org/10.31165/nk.2007.12.20

Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. 1995. A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM Journal on Scientific Computing* 16, 5 (Sept. 1995), 1190–1208. https://doi.org/10.1137/0916069

Neill D. F. Campbell and Jan Kautz. 2014. Learning a Manifold of Fonts. *ACM Trans. Graph.* 33, 4 (July 2014), 91:1–91:11. https://doi.org/10.1145/2601097.2601212

Barbara S. Chaparro, A. Dawn Shaikh, Alex Chaparro, and Edgar C. Merkle. 2010. Comparing the Legibility of Six ClearType Typefaces to Verdana and Times New Roman. *Information Design Journal* 18, 1 (Aug. 2010), 36–49. https://doi.org/10.1075/idj.18.1.04cha

Karen Cheng. 2020. *Designing type.* Yale University Press.

Saemi Choi, Toshihiko Yamasaki, and Kiyoharu Aizawa. 2016. Typeface Emotion Analysis for Communication on Mobile Messengers. In *Proceedings of the 1st International Workshop on Multimedia Alternate Realities (AltMM '16)*. Association for Computing Machinery, New York, NY, USA, 37–40. https://doi.org/10.1145/2983298.2983305

Jacob Cohen. 1994. The Earth Is Round ($p < .05$). *American Psychologist* 49, 12 (1994), 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

Stephen Coles. 2013. *The geometry of type: The anatomy of 100 essential typefaces.* Thames & Hudson.

Geoff Cumming. 2013. *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis.* Routledge.

Lucas de Groot. 2020. Interpolation Theory. https://www.lucasfonts.com/learn/interpolation-theory

Stanislas Dehaene. 2010. *Reading in the brain: The new science of how we read.* Penguin Group USA.

Matthew J. Denwood. 2016. Runjags: An R Package Providing Interface Utilities, Model Templates, Parallel Computing Methods and Additional Distributions for MCMC Models in JAGS. *Journal of Statistical Software* 71 (July 2016), 1–25. https://doi.org/10.18637/jss.v071.i09

Jonathan Dobres, Nadine Chahine, Bryan Reimer, David Gould, Bruce Mehler, and Joseph F. Coughlin. 2016. Utilising psychophysical techniques to investigate the effects of age, typeface design, size and display polarity on glance legibility. *Ergonomics* 59, 10 (2016), 1377–1391. https://doi.org/10.1080/00140139.2015.1137637 PMID: 26727912.

Frank H. Durgin and Zhi Li. 2011. The Perception of 2D Orientation Is Categorically Biased. *Journal of Vision* 11, 8 (July 2011), 13–13. https://doi.org/10.1167/11.8.13

Gustav Theodor Fechner. 1948. Elements of Psychophysics, 1860. In *Readings in the History of Psychology.* Appleton-Century-Crofts, East Norwalk, CT, US, 206–213. https://doi.org/10.1037/11304-026

Cristian Felix, Steven Franconeri, and Enrico Bertini. 2018. Taking Word Clouds Apart: An Empirical Investigation of the Design Space for Keyword Summaries. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (Jan. 2018), 657–666. https://doi.org/10.1109/TVCG.2017.2746018

Yue Gao, Yuan Guo, Zhouhui Lian, Yingmin Tang, and Jianguo Xiao. 2019. Artistic Glyph Image Synthesis via One-Stage Few-Shot Learning. *ACM Transactions on Graphics* 38, 6 (Nov. 2019), 185:1–185:12. https://doi.org/10.1145/3355089.3356574

Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 1995. *Bayesian Data Analysis.* Chapman and Hall/CRC, New York. https://doi.org/10.1201/9780429258411

Han Han and Miguel Nacenta. 2020. The Effect of Visual and Interactive Representations on Human Performance and Preference with Scalar Data Fields. In *Proceedings of the 46th Graphics Interface Conference* (Toronto, Ontario, Canada) *(GI '20)*. Canadian Human-Computer Communications Society. https://doi.org/10.20380/GI2020.23

Yannis Haralambous. 1993. Parametrization of PostScript fonts through METAFONT-an alternative to Adobe Multiple Master fonts. *Electronic Publishing* 6, 3 (Sept. 1993), 145–157. https://hal.archives-ouvertes.fr/hal-02100464

Yannis Haralambous. 2007. *Fonts & encodings.* O'Reilly Media, Sebastopol, Calif.

Yannis Haralambous and Tereza Haralambous. 2003. Characters, Glyphs and Beyond. In *Symposium on Glyphs, 21st Century COE Program.* Kyōto University, Kyoto, Japan. https://hal.archives-ouvertes.fr/hal-01290554

Lane Harrison, Fumeng Yang, Steven Franconeri, and Remco Chang. 2014. Ranking Visualizations of Correlation Using Weber's Law. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (Dec. 2014), 1943–1952. https://doi.org/10.1109/TVCG.2014.2346979

Tamir Hassan, Changyuan Hu, and Roger D. Hersch. 2010. Next Generation Typeface Representations: Revisiting Parametric Fonts. In *Proceedings of the 10th ACM Symposium on Document Engineering (DocEng '10)*. Association for Computing Machinery, New York, NY, USA, 181–184. https://doi.org/10.1145/1860559.1860596

Hideaki Hayashi, Kohtaro Abe, and Seiichi Uchida. 2019. GlyphGAN: Style-consistent Font Generation Based on Generative Adversarial Networks. *Knowledge-Based Systems* 186 (Dec. 2019), 104927. https://doi.org/10.1016/j.knosys.2019.104927

Jeffrey Heer and Michael Bostock. 2010. Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. ACM, New York, NY, USA, 203–212. https://doi.org/10.1145/1753326.1753357

Douglas R Hofstadter and Emmanuel Sander. 2013. *Surfaces and essences: Analogy as the fuel and fire of thinking.* Basic Books.

Keith Houston. 2013. *Shady Characters: The Secret Life Of Punctuation, Symbols And Other Typographical Marks.* WW Norton, New York ; London.

Catherine Q. Howe and Dale Purves. 2005. Natural-Scene Geometry Predicts the Perception of Angles and Line Orientation. *Proceedings of the National Academy of Sciences* 102, 4 (Jan. 2005), 1228–1233. https://doi.org/10.1073/pnas.0409311102

Changyuan Hu and Roger D. Hersch. 2004. Perceptually-Tuned Grayscale Characters Based on Parametrisable Component Fonts. In *Digital Documents: Systems and Principles*, Peter King and Ethan V. Munson (Eds.). Number 2023 in Lecture Notes in Computer Science. Springer Berlin Heidelberg, 69–76. http://link.springer.com/chapter/10.1007/978-3-540-39916-2_6

John Hudson. 2016. Introducing OpenType Variable Fonts. https://medium.com/@tiro/https-medium-com-tiro-introducing-opentype-variable-fonts-12ba6cd2369#.80oysl6l8

Sarah Hyndman. 2016. *Why fonts matter.* Random House.

Florian Kadner, Yannik Keller, and Constantin Rothkopf. 2021. AdaptiFont: Increasing Individuals' Reading Speed with a Generative Font Model and Bayesian Optimization. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Number 585. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3411764.3445140

Matthew Kay and Jeffrey Heer. 2016. Beyond Weber's Law: A Second Look at Ranking Visualizations of Correlation. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (Jan. 2016), 469–478. https://doi.org/10.1109/TVCG.2015.2467671

Stanley A. Klein. 2001. Measuring, Estimating, and Understanding the Psychometric Function: A Commentary. *Perception & Psychophysics* 63, 8 (Nov. 2001), 1421–1455. https://doi.org/10.3758/BF03194552

Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica Hamrick, Jason Grout, Sylvain Corlay, Paul Ivanov, Damián Avila, Safia Abdalla, Carol Willing, and Jupyter Development Team. 2016. Jupyter Notebooks – a Publishing Format for Reproducible Computational Workflows. *Positioning and Power in Academic Publishing: Players, Agents and Agendas* (2016), 87–90. https://doi.org/10.3233/978-1-61499-649-1-87

Donald E. Knuth. 1986. *The metafont book.* Addison-Wesley Reading, MA.

Leonid L. Kontsevich and Christopher W. Tyler. 1999. Bayesian Adaptive Estimation of Psychometric Slope and Threshold. *Vision Research* 39, 16 (Aug. 1999), 2729–2737. https://doi.org/10.1016/S0042-6989(98)00285-5

John Kruschke. 2014. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan.* Academic Press.

John K. Kruschke. 2012. Complete Example of Right Censoring in JAGS (with Rjags). http://doingbayesiandataanalysis.blogspot.com/2012/01/complete-example-of-right-censoring-in.html

Gordon E. Legge and Charles A. Bigelow. 2011. Does Print Size Matter for Reading? A Review of Findings from Vision Science and Typography. *Journal of Vision* 11, 5 (2011), 8–8. http://iovs.arvojournals.org/article.aspx?articleid=2191906

Zhouhui Lian, Bo Zhao, Xudong Chen, and Jianguo Xiao. 2018. EasyFont: A Style Learning-Based System to Easily Build Your Large-Scale Handwriting Fonts. *ACM Transactions on Graphics* 38, 1 (Dec. 2018), 6:1–6:18. https://doi.org/10.1145/3213767

Ellen Lupton. 2010. *Thinking with Type, Second Revised and Expanded Edition: A Critical Guide for Designers, Writers, Editors, and Students* (2nd revised edition ed.). Princeton Architectural Press.

Alan M. MacEachren. 2004. *How maps work: representation, visualization, and design.* Guilford Press.

D. M. MacKay. 1963. Psychophysics of Perceived Intensity: A Theoretical Basis for Fechner's and Stevens' Laws. *Science* 139, 3560 (1963), 1213–1216. https://www.jstor.org/stable/1710451

J. S. Mansfield, G. E. Legge, and M. C. Bane. 1996. Psychophysics of Reading. XV: Font Effects in Normal and Low Vision. *Investigative Ophthalmology & Visual Science* 37, 8

(July 1996), 1492–1501. https://iovs.arvojournals.org/article.aspx?articleid=2161324
Constant Manteau, Miguel Nacenta, and Michael Mauderer. 2017. Reading Small Scalar Data Fields: Color Scales vs. Detail on Demand vs. FatFonts. In *Proceedings of the 43rd Graphics Interface Conference (GI '17)*. Canadian Human-Computer Communications Society, Waterloo, CAN, 50–56. https://doi.org/10.20380/GI2017.07
Yannick Mathey and Prototypo Team. 2020. *Create Unlimited Type Variations with a Few Swipes.* https://www.prototypo.io/
Clyde D McQueen III and Raymond G Beausoleil. 1993. Infinifont: A Parametric Font Generation System. 6, 3 (1993), 117–132.
Microsoft. 2018. *OpenType Specification v1.8.3.* Microsoft. https://docs.microsoft.com/en-us/typography/opentype/spec/
Tomo Miyazaki, Tatsunori Tsuchiya, Yoshihiro Sugaya, Shinichiro Omachi, Masakazu Iwamura, Seiichi Uchida, and Koichi Kise. 2020. Automatic Generation of Typographic Font From Small Font Subset. *IEEE Computer Graphics and Applications* 40, 1 (Jan. 2020), 99–111. https://doi.org/10.1109/MCG.2019.2931431
Miguel Nacenta, Uta Hinrichs, and Sheelagh Carpendale. 2012. FatFonts: Combining the Symbolic and Visual Aspects of Numbers. In *Proceedings of the International Working Conference on Advanced Visual Interfaces* (Capri Island, Italy) *(AVI '12)*. Association for Computing Machinery, New York, NY, USA, 407–414. https://doi.org/10.1145/2254556.2254636
Roel Nieskens. 2016. Variable Fonts: the Future of (Web) Type. https://typographica.org/on-typography/variable-fonts/ Library Catalog: typographica.org.
Yoshi Ohno. 2000. CIE Fundamentals for Color Measurements. *NIP & Digital Fabrication Conference* 2000, 2 (Jan. 2000), 540–545.
H. Q. Phan, H. Fu, and A. B. Chan. 2015. FlexyFont: Learning Transferring Rules for Flexible Typeface Synthesis. *Computer Graphics Forum* 34, 7 (2015), 245–256. https://doi.org/10.1111/cgf.12763
Martyn Plummer et al. 2003. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing* (Vienna, Austria.), Vol. 124. 10.
Keith Rayner, Alexander Pollatsek, Jane Ashby, and Charles Clifton Jr. 2012. *Psychology of reading.* Psychology Press.
Luz Rello, Martin Pielot, and Mari-Carmen Marcos. 2016. Make It Big! The Effect of Font Size and Line Spacing on Online Readability. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 3637–3648. https://doi.org/10.1145/2858036.2858204
Thomas Sanocki. 1987. Visual Knowledge Underlying Letter Perception: Font-specific, Schematic Tuning. *Journal of Experimental Psychology: Human Perception and Performance* 13, 2 (1987), 267–278. https://doi.org/10.1037/0096-1523.13.2.267
Thomas Sanocki. 1988. Font Regularity Constraints on the Process of Letter Recognition. *Journal of Experimental Psychology: Human Perception and Performance* 14, 3 (1988), 472. http://psycnet.apa.org/journals/xhp/14/3/472/
Ariel Shamir and Ari Rappoport. 1998. Feature-Based Design of Fonts Using Constraints. In *Electronic Publishing, Artistic Imaging, and Digital Typography*, Roger D. Hersch, Jacques André, and Heather Brown (Eds.). Number 1375 in Lecture Notes in Computer Science. Springer Berlin Heidelberg, 93–108. http://link.springer.com/chapter/10.1007/BFb0053265
Nick Sherman. 2021. *Variable Fonts Web Site.* https://v-fonts.com/
Yuto Shinahara, Takuro Karamatsu, Daisuke Harada, Kota Yamaguchi, and Seiichi Uchida. 2019. Serif or Sans: Visual Font Analytics on Book Covers and Online Advertisements. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*. 1041–1046. https://doi.org/10.1109/ICDAR.2019.00170
Barbara G. Shortridge. 1979. Map Reader Discrimination of Lettering Size. *The American Cartographer* 6, 1 (Jan. 1979), 13–20. https://doi.org/10.1559/152304079784022781 Publisher: Taylor & Francis.
Barbara G. Shortridge and Robert B. Welch. 1982. The Effect of Stimulus Redundancy on the Discrimination of Town Size on Maps. *The American Cartographer* 9, 1 (Jan. 1982), 69–78. https://doi.org/10.1559/152304082783948277 Publisher: Taylor & Francis.
Hendrik Strobelt, Daniela Oelke, Bum Chul Kwon, Tobias Schreck, and Hanspeter Pfister. 2016. Guidelines for Effective Usage of Text Highlighting Techniques. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (2016), 489–498. https://doi.org/10.1109/TVCG.2015.2467759
Miles A. Tinker. 1944. Criteria for Determining the Readability of Type Faces. *Journal of Educational Psychology* 35, 7 (Oct. 1944), 385–396. https://doi.org/10.1037/h0055211
Seiichi Uchida, Yuji Egashira, and Kota Sato. 2015. Exploring the World of Fonts for Discovering the Most Standard Fonts and the Missing Fonts. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. 441–445. https://doi.org/10.1109/ICDAR.2015.7333800
Yizhi Wang, Yue Gao, and Zhouhui Lian. 2020. Attribute2Font: Creating Fonts You Want from Attributes. *ACM Transactions on Graphics* 39, 4 (July 2020), 69:69:1–69:69:15. https://doi.org/10.1145/3386569.3392456
Wikipedia contributors. 2020. Em (typography). https://en.wikipedia.org/w/index.php?title=Em_(typography)&oldid=984185316 Page Version ID: 984185316.
Andy T. Woods, Carlos Velasco, Carmel A. Levitan, Xiaoang Wan, and Charles Spence. 2015. Conducting Perception Research over the Internet: A Tutorial Review. *PeerJ* 3 (July 2015), e1058. https://doi.org/10.7717/peerj.1058
Chang Xiao, Cheng Zhang, and Changxi Zheng. 2018. FontCode: Embedding Information in Text Documents Using Glyph Perturbation. *ACM Transactions on Graphics* 37, 2 (Feb. 2018), 15:1–15:16. https://doi.org/10.1145/3152823

A PARAMETRIC FONT PARAMETER DETAILS

A.1 Further Detail on Selection of Parameters

Following on the criteria for the selection of parameters introduced in Section 3, (a) and (d) support compatibility with type design practices and terminologies, (b) makes the study feasible, focused on letter shapes, and novel (general visual parameters such as color and size[13] have been studied elsewhere, e.g. [Alexander et al. 2018; Bateman et al. 2008; Shortridge 1979; Strobelt et al. 2016]), and (c) and (d) assure that they can be widely used for infotypographic purposes, including to represent multiple interval and ratio data values simultaneously. The authors agreed on the final selection by iteratively refining the implementation while considering literature in font morphology (e.g., [Bauermeister 1988], [Haralambous 2007, p.367]), existing variable and parametric systems (e.g., [Knuth 1986; Mathey and Prototypo Team 2020]) and research on font use for information display (e.g., [Brath and Banissi 2014b]).

The Prototypo platform provided good support for the implementation of these parameters. One of the main advantages of the John Fell variable font was that many existing parameters were implemented with simple linear relationships between levels and geometrical shapes. Other existing parametric/variable fonts might be programmed with other relationships (e.g., quadratic).

A.2 Range Selection

Regarding the criteria for choosing the ranges and neutral points introduced in Section 3, we enforced (a)—practicality and readability of the extremes—by applying our own judgement and expertise, and (b)—avoidance of visual artifacts created by the implementation—through an iterative process in which we plotted all possible variants of fonts for three values in each parameter (minimum, neutral and maximum—a total of $3^7 = 2187$ variants) and inspected them one by one for unsightly artefacts and difficult to read and distinguish letters. When either were found, we reduced the parameter range until the artifact or confusion disappeared or, in a few cases, we were able to change the font implementation to fix the problem. Some of the generated inspection pdfs for the process are included in the supplementary materials. The neutral points were chosen so that the completely neutral font (all parameters in neutral values) had a familiar appearance. For some parameters such as *slant*, this is straightforward (vertical font, 0 degrees), but for others such as aperture and weight we had to use our own subjective criteria and reference common types such as Times New Roman. At the end, the process carried out to enforce (b) was much more limiting in the choice of ranges than (a).

[13] Point size or optical size of a font do change shape, but this typically translates into modifications of other aspects of the letter shape, such as weight.

B PARTICIPANT SELECTION AND FILTERING

Participants were from one of six countries with English as a dominant official language (UK, US, Ireland, Australia, Canada, NZ). Participant compensation was based on UK minimum wage at the time (£6.80 per hour) pro-rated by the approximate time to complete the experiment that we obtained through an informal empirical assessment based on pilot studies (times and rates were different for each experiment).

Only data from participants who completed all the tasks for a given parameter were included in the analysis. We filtered the data of participants who were deemed to be "random clickers", implausibly fast, etc. This filtering was based on a battery of measurements such as the range of their answers (narrow range indicates a participant clicking on the same answer regardless of stimuli), the number of times the answer was provided faster than 1 second (indicates compromising accuracy to finish too fast), or the normalized deviation with respect to the overall set of participants in the correlation between input and output (outliers that show unusually low correlations are likely to be answering randomly). We eliminated participants who failed more than a certain number of these tests (tests and acceptance thresholds are different for each study and reported below in Sections D.1 and E.1).

C GENERAL ELEMENTS AND IMPLEMENTATION OF THE BAYESIAN ANALYSIS

C.1 Justification of Bayesian Analysis

We chose a contemporary Bayesian analysis framework because of four main reasons: 1) it provides a consistent approach for the comparison of models; 2) it enables a coherent approach to the analysis of truncated data; 3) it makes explicit our modeling choices; and 4) it facilitates the interpretation of the results. We do not discuss here the comparison between Bayesian, Frequentist and other approaches to statistics (see, e.g.,Refs. [Cohen 1994; Cumming 2013; Kruschke 2014]). However, despite the intensive computation cost incurred due to the non-analytical MCMC approach and the need to write custom analysis in Python and R, we are doubtful that a comparably sophisticated analysis would have been possible to us with a different statistical framework.

One of the drawbacks of the Bayesian approach is that it requires a significant amount of analysis decisions, and some adjustments to make MCMC chains converge and cover the parameter space. These are tedious and space-consuming, but to provide readers with sufficient detail to assess reliability and validity we provide: a) the entirety of the analysis and the output from the simulations in the Supplementary Materials; and, b) a detailed account of the main analysis decisions in the Appendices.

C.2 MCMC General Parameters and Implementation

In all MCMC simulations of our analysis we attained a good mix of chains and effective sample sizes (ESS) of above 10,000 for the parameters of interest. The particular assumptions and models are different for each experiment and described in their corresponding sections. We used python scripts within Jupyter Notebooks [Kluyver et al. 2016] for conditioning the data (running Python 3.6.8), and R (v.3.6.0) within Jupyter Notebooks for the rest of the analysis. The MCMC simulations were created through JAGS [Plummer et al. 2003] (v.4.2.0), which was interfaced through the *runjags* package [Denwood 2016]. In general, prior functions were made suitably uninformative to avoid biasing the data. This is consistent with the lack of accurate information from prior work as to the perceptual characteristics of fonts. Generally, prior functions were Gaussian in shape and with very large variances (multiple times larger than the final variances of the posterior probabilities that we found). For some parameters it is necessary to truncate the possible values of the prior because certain combinations of values fail to evaluate the mathematical functions discussed in Section D.2. All these parameters were carefully adjusted by hand, with care to avoid any type of bias. All priors are available for inspection in the analysis scripts in the Supplementary Materials.

We considered using Student-t robust distributions to fit the models instead of the usual Normal (Gaussian) distribution. The Student-t distribution is a generalization of the Normal distribution that can account for heavier tails, and contains an additional parameter. However, we found that the additional parameter of the distribution made many of the model chains unstable and heterogeneous. Therefore we use Normal distributions by default.

C.3 Censoring or Truncating

Because the data collected is necessarily constrained to a bounded interval (e.g., participant responses are bound to the interval -0.5 to 0.5 in Experiment 1), the fitting of the models needs to take into account that sample measures at the ends of these scales might be representative of values behind the limit of the scale. Fortunately, this unavoidable issue can be dealt gracefully through the Bayesian framework and, in particular, through JAGS by *censoring* (see [Kruschke 2012]). Essentially, we indicate to the MCMC algorithm when a measurement has been censored (for us, these are the measurements at the edges of the scale in Experiment 1), and then the model is estimated taking into account that those measurements might be representative of samples that would have been anywhere beyond the measurable scale. We apply censoring in both experiments, which allows us to use Normal curves to estimate the parameters in our models. We also considered, instead, truncating the model function at the limits of the scale; this does not seem to be the recommended procedure in the Bayesian literature. We set up a small experiment to compare the effect of truncating versus censoring with data and we found that truncation of the model underestimated the parameters of the model (compared with truncating) on an upper-censored sample set, hence we used censoring. The data is pre-processed for censoring using the function `prepareCensoredData` in either of the main analysis files included in the Supplementary Materials (`Exp1_MCMC_analysis.ipynb` and `Exp2_MCMC_analysis.ipynb`).

D EXPERIMENT 1 ANALYSIS

D.1 Data Filtering and Participants

To avoid noise from participants who mindlessly clicked through the trials we established a battery of tests on the data. This is common practice in crowdsourced experiments (see, e.g., [Heer and Bostock 2010]). Due to the nature of our experiment it is not practical to use

simple "check tests" with trivial answers, since these would be too easy to identify by rogue participants.

The battery of tests check whether too many of the stimuli at the the scale's ends were judged far from their end, have very low correlation between the stimuli and the responses, a high proportion of middle value responses, unfeasibly short trial times, and answer ranges that were too narrow. The check's parameters were conservative (err on the side of including noisy participants). This is because we believe that it is safer to underestimate the amount of noise of the perceptual measurements by including rogue participants that add noise to our measurements, hence *underestimating* people's ability to see differences than the opposite (overestimating the general ability of people to perceive the parameters by selecting only the better performing participants). A data set was eliminated if it failed two or more of the tests listed above. The exact criteria is available from the code in the Supplementary Materials, in the `EXP1_Read_and_Clean.ipynb` analysis file.

The 40 datasets filtered did not comply with two or more of the pre-established criteria. Some of our criteria depend on the averages or distributions in the group of participants completing this parameter, which explains why there is not a uniform number of data sets per parameter. The distributions of filtered sets per parameter are listed in Table 13. Participants received a compensation of £2.38 for each set. The median time to complete a single trial was 4.5 seconds. Only two participants had experience designing fonts.

Fig. 13. Participant data set numbers for Experiment 1.

	weight	width	contrast	xHeight	slant	serifs	aperture	total
Completed	36	35	36	38	35	35	39	254
Compliant	30	31	30	30	32	31	30	214
Priors	10	11	10	10	11	11	10	70
Main	20	20	20	20	21	20	20	140

D.2 Model Selection and Parameterization

To cover a wide range of behaviours of the perception curve we tested four model functions (equations 2–5 in Section 6.4.1).

Equations 2-4 have 3 parameters, whereas Equation 5 has four. In standard statistical fitting of the curves (e.g., maximizing R^2) this would lead to the 4-parameter model always being more accurate than the others, since it can, by definition, accommodate more variation. Standard solutions to this problem include calculating of alternative measures of fit, such as the Akaike information criterion [1974], which penalize models with more parameters. The Bayesian model comparison procedure naturally compensates for this without having to penalize the models explicitly. For an explanation of the principle behind this see Ref. [Kruschke 2014, p.291].

The parameterizations are chosen for interpretability and to be comparable to each other. For example, in the sigmoid model x_{so} determines the horizontal position of the 's'-shaped curve, higher values of r_s increase the slope of the curve in its middle, and κ_s adjusts the vertical range of the codomain of the function. Figure 5 displays one example of each of the models with a representative choice of parameter values. We do not claim to be able to differentiate precisely between specific variations of the perception curve, e.g., concave quadratic versus exponential; this might be useful for perception theory—pointing to the possible underlying neural mechanisms—but would require a different experiment design and more data. Instead we we characterize the general shape.

The quadratic and cubic models can describe perceptual biases where the perceived value differences between two stimuli at the same input distance (in linear terms) decreases (or increases) with higher levels of the value (e.g., consistent with the Fechner-Weber's law [1948]). E.g., a thick letter with stems 1mm wide is less perceptually different from a letter 0.1mm thicker than a thin letter of 0.1mm from the equally thicker letter of 0.2mm width stems. Sigmoid (and inverse-sigmoid) curves approximate biases where perception differentiates better range's center (respectively, the edges).

D.3 MCMC Simulation Chain Procedures

To calculate the fitting parameters and compare the probabilities of the different models we use a series of MCMC simulations in six phases. The three main desired outcomes of this analysis process are consistent with the research questions in the paper: first, compare the different models (functions) in Subsection 6.4.1, which answers **Q2**; second, fit the parameters of the best model, including the variability (noise parameter) to answer **Q1**, and third, refine the best model to account for the variability of individuals (**Q4**). To compare models we calculate Bayes factors [Kruschke 2014], which are, in essence, the probabilities of individual *states*, given the experimental data, obtained from a single simulation in which each model appears as a single discrete state.

D.4 Priors, Pseudos and Multi-Phase Computations for Model Selection (Exp 1)

In Experiment 1, we use the data to ascertain the perceptual response's shape for each parameter. We assume no prior knowledge of the probabilities of different models; we considered all models as equally likely *a priori*. Because the choice of priors can influence Bayes Factors probabilities, we follow Kruschke's recommended procedure: obtaining a refined set of priors from a subset of the data (for each model) before running the model comparison [2014, p.294]. This minimizes the influence of the choice of priors because the naïve priors are easily overwhelmed by even small amounts of data. In our case, we select the data from approximately one third of participants for this purpose in each of the typographic parameter subanalyses. In the following text we refer to the data coming from these participants as $data_{priors}$, the data from participants not in this group as $data_{main}$, and all the data as $data_{all}$. For model selection we use data were we average values per cell, i.e., we average all responses from the same participant for each of the levels of input. The latest stages of analysis, once the model has been chosen, use instead the individual data points per participant (unaveraged). References to averaged data are appended with "avgd" on the subindex (e.g., $data_{priors_{avgd}}$, $\overline{data_{all_{avgd}}}$).

The following phases are run for each of the typographic parameters independently:

Phase 1. Run separate simulations of each model with naïve priors and $data_{priors_{avgd}}$ as main data. The parameter posteriors serve as fine-tuned priors for the parameters in the next phase.

Phase 2. Run separate simulations of each model with the fine-tuned priors and $data_{main_{avgd}}$ as main data. The posterior of this process serves as *pseudo values* for the parameters of the next phase. Pseudo values are required to enable efficient running of multi-model comparisons in our MCMC setup.

Phase 3. Run simulations with all models, using priors from Phase 1, pseudo values from Phase 2 and $data_{main_{avgd}}$ as main data. Priors of the *states* (the prior probabilities of each model) are iteratively adjusted to achieve sufficient mix of all models.

Phase 4. Run one simulation with all models with the state priors calculated in Phase 3 and $data_{main_{avgd}}$ as main data. The results, when weighted with the priors [Kruschke 2014, p.279] provide the probabilities of each of the models.

Phase 5. Run only the winner model from Phase 4 by itself with $data_{priors_{avgd}}$ as priors and $data_{main_{avgd}}$ as main data. The most likely combination of joint parameter posteriors from this run are considered the best (canonical) model.[14]

Phase 6. Assuming the winner model from Phase 4, run the sophisticated version of the model in which where each participant has its own variance multiplier, and variance is modeled separately for each of the levels of the input values (see Equation 6). This requires use of the non-averaged data ($data_{priors}$ as priors and $data_{main}$ as main data).

D.5 Optimal Stimuli Analysis

In Section 7.3 and the tetragrams at the bottom of Figure 6's subfigures we use the calculated model to determine where in the scale we could optimally place discrete levels to minimize confusion in a 2, 3, 5 and 7 level scale. The procedure assumes Normal distribution of the noise around the expected mean estimated value from the data (using the most likely model). The variance is linearly interpolated at the continuous range of levels from the estimated variances at the measured levels in the most likely model. A numeric approximation algorithm based on Byrd et al.'s [1995] minimizes the probability of one class being perceived as part of another class. We also calculate the optimal decision boundaries between levels (displayed as vertical lines in the tetragram). These optimal levels are based on the assumption that all classes are equally likely. If the probabilities are not uniform, the algorithm can be modified to account for the imbalance, which will shift the optimal levels. The implementation of this optimization is in function `findOptimalCentres` of the `Exp1_MCMC_analysis.ipynb` file. The exact level locations calculated appear in a spreadsheet named `Exp1_results_splits_with_converted_values.xlsx` in the Supplementary Materials. The values of the optimal levels are provided both in normalized and denormalized scales.

[14] We use an iterative calculation of the most densely populated n-dimensional hypercube of the parameter space (where n is the number of parameters for that model—3 or 4 in the case of the cubic model), although the differences with independently taking medians of each parameter are very small.

D.6 Additional Result Materials

D.6.1 Bayes Factors. The decisions regarding which is the best model for each parameter are made based on Bayes Factors, which indicate the probability of a given model with respect to the other models involved in the comparison. The best model is simply the most likely. However, it can be informative to see the probability that the MCMC algorithm assigns to the different models for each parameter. Table 4 presents that table. Some parameters show much more clear decisions than others, but the results were stable across multiple runs of the algorithm.

Table 4. Bayes Factors for Experiment 1. Values (in %) indicate probability of a given model being the underlying model that produces the data against the other alternatives (for each typographic parameter).

	Quadratic	Sigmoid	Inv. Sigm.	Cubic
Slant	9.6	16.7	43.6	30.1
Weight	56.2	<0.1	0.2	43.6
Width	65.0	4.0	<0.1	31.0
xHeight	2.9	97.1	<0.1	<0.1
Contrast	43.0	1.5	<0.1	54.3
Aperture	17.3	50.8	27.2	4.7
Serifs	38.2	40.0	1.3	20.6

E EXPERIMENT 2 ANALYSIS

E.1 Data Filtering and Participants

Data filtering for Experiment 2 is similar to that of Experiment 1 (described in section D.1). The main difference is that we tested a much larger set of constraints (e.g., a high probability of random clicks, or always clicking on the same side), too long to list here (see details in the `EXP2_Read_and_Clean.ipynb` script). Because the battery of tests is much larger, the threshold of number of tests to fail is larger (9 or more).

48 data sets were excluded because they did not comply with 9 or more of the pre-established filtering criteria. Participants received a compensation of between £2.04 and £2.72 for each set, based on a duration of study approximated through pilot testing (24 minutes for aperture and slants, 18 minutes for weight, 21 minutes for the rest). The median time to complete a single trial was 3.38 seconds. Only one participant had experience designing fonts.

Fig. 14. Participant data set numbers for Experiment 2.

	weight	width	contrast	xHeight	slantL	slantR	serifs	aperture	total
Completed	30	31	34	33	33	35	36	53	285
Compliant	30	31	30	32	32	31	31	20	237

E.2 Estimation of JND Threshold and Slope

Experiment 2 is based on a *two alternative forced choice* paradigm. At each step of a staircase, the participant has to decide which of two rendered words is the one with the higher value of a particular parameter. Which rendering appears to the left or to the right is random, but the distance in the parameter space between the stimuli varies so that the threshold at which participants cannot see a difference can be found. The algorithm that determines which stimuli are shown to the participant and in which order is adaptive, which means that it chooses different stimuli depending on previous answers. A variety of staircase procedures have been proposed; most try to minimize the number of steps that it takes to reach an stable measurement of the JND. For example, typical staircase procedures start by showing rendering pairs that are very different to each other, and progressively reduces this difference until it approximates the point at which the participant is simply guessing.

We use a staircase procedure proposed by Kontsevich et al., which uses a calculation of entropy at each step to decide which distance between stimuli is most likely to add new information to the calculation of the JND [Kontsevich and Tyler 1999]. We use an implementation created by Joseph J. Glavan which is distributed with a GNU license and included in the Supplementary Materials in the `./analysis/exp2_JND/adaptive/psy.py` script. This implementation is used from the `StaircaseProcedure.py` script, which also contains the exact parameters of the staircase configuration.

The procedure output is, for each staircase, the estimated level of the parameter where the participant's probability of making a mistake is 25%. This is what we use as an estimation of the Just Noticeable Difference (JND). This is a common choice in psychophysical studies, since it is midway from random performance (50% chance of being correct), and perfect performance (100% chance of being correct). This JND value is the level at which the psychometric function (probability of successful discrimination as a function of the magnitude of the difference between presented stimuli) crosses the 75% probability threshold. This function, which is often modeled as a sigmoid, can also vary in shape; i.e., the transition between low differentiability and high differentiability around the level that crosses the 75% threshold can be steep, or flat. This is what is referred to as the *slope* calculation.

Our procedure also enables calculation of the slope, which is a secondary output (besides the JND). The slope outputs are included in the data. However, we used staircases with 35 steps which turned out to be not sufficient to provide an accurate enough estimation of the slope. An estimated sufficient number of steps to provide a reasonable estimation of the slope is 300, which was out of the scope for testing in this experiment. Since the slope measurement was too noisy we do not report results about the slope.

E.3 Modeling

In Experiment 2 the main input data for the analysis is the estimated Just Noticeable Difference (JND) of the procedure, as described above. The main analysis aims to model how the JND varies along the scale of each paramters. As described in Section 8.4 of the paper, we did not try to do a model comparison, since the five levels at which we measure, and the added complication of levels that are approached from above and below would probably have made the often subtle distinctions between models difficult.

E.4 Truncating vs. Censoring in Experiment 2

As in Experiment 1, the nature of the measured data points requires that we apply censoring (see Section C.3). The thresholds at which to censor for Experiment 2 are, however, harder to determine because they depend on the staircase procedure. There are two different censoring levels in this experiment. On the upper end, we use censoring to indicate to the MCMC procedures that JND values above a certain level are not to be considered as estimations of the shape of the Normal distribution, since they are the result of the process of random answers to the staircase decisions, which would introduce bias to the fitting of the LogNormal noise distribution. The location of the threshold is also influenced by the magnitude of the maximum difference that a staircase can span (e.g., a staircase for the largest level—+0.5—approaching from below can display differences of up to 1.0 in parameter range magnitude, whereas a downward-approaching staircase for the second to highest level maximum stimuli difference magnitude is 0.25 units.

We therefore calculate upper thresholds of the JND estimations by simulating many random staircases and averaging the JNDs produced. The censoring procedure uses those thresholds to censor any data that is larger. This is done separately for each staircase.

Censoring levels are in `./analysis/calculated_parameters/` of the Supplementary Materials (in Python pickle format). The calculation procedure is in the `EXP2_MCMC_Read_and_Clean.ipynb` Python script in the `./analysis/exp2_JND/` folder.

Censoring is also necessary for the lower bounds because participants might be able to discern differences that go beyond the granularity available in our configuration of the font. Not censoring these values at the lower level would also skew the noise distribution fitting, since the LogNormal distribution of the noise makes it particularly sensitive at low levels.

The censoring levels that we calculated are displayed in Figure 11 with hollow triangles. Downward pointing triangles indicate upper censoring thresholds and upward facing triangles indicate lower censoring thresholds. When small numbers appear—for each of the five levels—at the bottom (for each of the parameters) or top of the scale, these indicate the number of datapoints that were censored (no numbers means no censoring). These numbers are generally small in all parameters. Some upper thresholds are not indicated in the figure because they are above the vertical scale of the figure (i.e., larger than 0.5 units in each parameter). Red triangles are for staircases coming from above and blue triangles are for staircases coming from below.

E.5 Optimal Stimuli Analysis

To estimate reasonable locations for levels of the parameters in a discrete scale we calculated sets of levels without overlapping estimated JNDs. The estimated JND is calculated using the fitted cubic function. The algorithm also has to take into account that each level has JNDs that approach from above and below and might be different, which means that the best set of levels might be calculated by starting from

the top, going down or from the bottom, going up. The full algorithm is in the `findBestRangeOfNonOverlappingJND()` function of the `Multi-parameter-results-and-graphs.ipynb` file in the `./analysis/exp2_JND/` folder of the Supplementary Materials.

Unlike in Experiment 1 we cannot estimate locations of levels that would result in different levels of error. This would have been possible with a reliable estimation of the slope of the psychometric function, which we could not get (see Section E.2 above).

F PARAMETERS OF REPRESENTATIVE AND PARAMETRIC FONTS

`./additional_materials/parameter_values.pdf` in the Supplementary materials contains a table that characterizes common and relevant fonts in terms of the infotypographic parameters that we define in Section 3 of the paper.