Authors version as accepted by journal
Appears as:
Alexander, E. C., Chang, C.-C., Shimabukuro, M., Franconeri, S., Collins, C., & Gleicher, M. (2018). Perceptual Biases in Font Size as a Data Encoding. IEEE Transactions on Visualization and Computer Graphics, 24(8), 2397–2410. http://doi.org/10.1109/TVCG.2017.2723397

Perceptual Biases in Font Size as a Data Encoding

Eric Alexander, Chih-Ching Chang, Mariana Shimabukuro,
Steven Franconeri, Christopher Collins, *Member, IEEE,* and Michael Gleicher, *Member, IEEE*

Abstract—Many visualizations, including word clouds, cartographic labels, and word trees, encode data within the sizes of fonts. While font size can be an intuitive dimension for the viewer, using it as an encoding can introduce factors that may bias the perception of the underlying values. Viewers might conflate the size of a word's font with a word's length, the number of letters it contains, or with the larger or smaller heights of particular characters ('o' vs. 'p' vs. 'b'). We present a collection of empirical studies showing that such factors—which are irrelevant to the encoded values—can indeed influence comparative judgements of font size, though less than conventional wisdom might suggest. We highlight the largest potential biases, and describe a strategy to mitigate them.

Index Terms—Text and document data, cognitive and perceptual skill, quantitative evaluation.

1 Introduction

With the growing amount of textual data available to researchers, methods of visualizing such data are of increasing importance. Text visualizations support analysts in many tasks, including forming a gist of a collection of documents, seeing temporal trends, and finding important documents to read in detail. One common method for encoding data using text rendering is to vary the font size. The importance and impact of font size as an encoding can be seen in a wide variety of contexts, from word cloud applications [1], [2], [3], to cartographic labeling [4], [5], to a number of different hierarchical visualization tools [6], [7].

However, there has been some question of how effective people are at judging font size encodings [8]. Such concerns arise in part because there are many ways in which words vary with one another outside of font size. In particular, two words with the same font size can vary tremendously in their *shape*. Longer words with more letters take up more area on the screen. The glyphs for some letters are inherently taller or wider than others. Kerning and tracking can create diverse spacing between characters. Differences in font would exacerbate these problems, but even the same font can be rendered differently depending on the platform. Other potential factors that could skew perception include color, font weight, and a word's semantic meaning [1], [3], [9], [10], [11].

We are interested in better understanding the ways in which these factors may bias font size perception. Such an understanding is important for knowing how much we can trust interpretations of data based on font size encodings. Measuring potential biases can also give us a way of finding limits on the kinds of tasks for which font size can be used—and seeing whether or not there are ways in which those limits can be stretched. Additionally, we can begin to tease apart the mechanisms that create those limits in a way that may inform the use of similar methods in different contexts.

In this paper, we focus specifically on the degree to which a **word's shape** can affect impressions of its font size. We present the results from a series of crowdsourced experiments in which participants were asked to judge font size within word cloud visualizations. In each experiment, we varied the words along one of the axes described above. We found that, in general, performance was high enough to call into question some existing notions of the limits of the encoding. However, there were conditions in which participants' perception of font size was biased. In particular, in cases where some physical attribute of the word, such as width, *disagreed* with its font size, accuracy dropped dramatically for many participants.

Fortunately, this effect can be corrected for. We describe a proof-of-concept method for debiasing font size encodings that uses colored tags sized proportionally to the data. We empirically show that our debiasing efforts improve performance even in the most pathological cases.

The main contributions of this paper are:

- An evaluation of user accuracy when making comparative judgements of font size encoding within a visualization, indicating that users may be better at making such judgements than conventional wisdom would suggest.
- A description of situations in which these judgements can be biased by attributes of the words being shown.
- A proof-of-concept method for debiasing visualizations in these situations using padded bounding boxes.

2 Related Work

Font size has been used to encode data across a number of visualization types, and to support a variety of tasks. Investigations of font size encoding have been largely focused on word clouds and their overall effectiveness, whereas our work focuses on the perceptual task of comparing word sizes under a variety of real-world conditions.

- *E. Alexander is with Carleton College, Northfield, MN, 55057. E-mail: ealexander@carleton.edu.*
- *C. Chang and M. Gleicher are with the University of Wisconsin-Madison, Madison, WI, 53706. E-mail: [chih-ching, gleicher]@cs.wisc.edu.*
- *M. Shimabukuro and C. Collins are with the University of Ontario Institute of Technology, Oshawa, Ontario, L1H 7K4, Canada. E-mail: [marianaakemi.shimabukuro, christopher.collins]@uoit.ca.*
- *S. Franconeri is with Northwestern University, Evanston, IL, 60208. E-mail: franconeri@northwestern.edu.*

Manuscript received April 19, 2005; revised August 26, 2015.

Fig. 1. To test whether attributes of words can affect perception of their font size, we highlighted words within word clouds and asked participants to choose the larger font. On the left, "zoo" has the larger font, but the length of "moreover" can bias participants toward choosing it as larger. On the right, "source" has the larger font, but the taller ascending and descending parts of "begged" can bias participants toward choosing it as larger.

The most familiar visualizations using font size encoding are tag clouds, more generally called word clouds. Word clouds represent variables of interest (such as popularity) in the visual appearance of the keywords themselves—using text properties such as font size, weight, or color [9]. One particularly popular example of word clouds is Wordle, an online tool for creating word clouds that encode word frequency information using font size [3]. Taking a cue from the popularity of word clouds, the Word Tree, which is an interactive form of the keyword-in-context technique, uses font size to represent the number of times a word or phrase appears [7].

Font size has also been used to encode data in cartographic visualizations, in typographic and knowledge maps. A typographic map represents streets using textual labels for street names while encoding spatial data such as traffic density, crime rate, or demographic data into the font size [4], [12]. In contrast, Skupin uses font size to indicate semantic clustering, adding a semantic hierarchy to his knowledge maps [5].

Rivadeneira et al. performed two experiments on word cloud effectiveness [11]. In the first, the effects of font size, location, and proximity to the largest word were investigated. The experiment results showed an effect of font size and position (upper-left quadrant) on recall; meanwhile, proximity showed no effect. In the second experiment, the authors evaluated impression formation and memory by varying font size and layout (e.g., alphabetical sorting, frequency sorting) of words in the cloud. Font size had a significant effect on recognition, but layout did not. However, the authors found that layout affected the accuracy of impression formation. From this evaluation, the authors concluded that word clouds are helpful for people to get a high-level understanding of the data, and for casual exploration without a specific target or goal in mind.

A study by Bateman et al. investigated the visual influence of word cloud visual properties (font size, tag area, tag width, font weight, number of characters, color, intensity and number of pixels) for the task of selecting the 10 "most important tags" [9]. Participants were asked to find the most attention-grabbing word out of a word cloud. They report that the features exerting the greatest visual influence on word clouds were font size, font weight, saturation and color. However, the authors did not look at user ability to accurately read data encoded with these features.

A study by Lohmann et al. [10] supports Bateman et al. [9] and Rivadeneira et al. [11] by reporting that words with larger font sizes attract more attention and are easier to find. However, none of these studies identify the magnitude of this effect for real-world use, or strategies for mitigating the biases. This knowledge is relevant because when encoding data into font size [4], [5], [7], [13] there is expectation from designers that people can perceive the difference in size to correctly understand the encoded data.

3 Experimental Task

There are many different documented tasks for which font size encodings have been used. These tasks include:

- **Gist-forming**: discerning the general meaning of a collection of words, taking their relative importance as coded by their font size into account [1], [11], [14].
- **Summary comparison**: making sense of juxtaposed sets of words from different sources [15], [16].
- **Word search**: finding a particular word in a visualization [9], [10], [11].
- **Retention**: being able to recall a word from a particular visualization, and to distinguish it from others [11].
- **Value reading**: reading a specific numerical value associated with text [13].
- **Order reading**: comparing words to determine relative value [9], [11].

It has been shown that font size encodings are not the proper design choice for a number of these tasks, most notably searching and retention, where simple ordering can be much more effective [11]. In general, font size encodings are more frequently used for subjective, high-level tasks like gist-forming. However, it is difficult to measure perceptual limitations with these tasks. For this study, we were not interested in measuring participants' cognitive ability to draw connections between groups of words, but rather in better understanding their *perceptual* abilities.

As such, in selecting a task for our experiments, we chose one that we believed would isolate the primitive *sub-task* of discerning information represented in font size. Specifically, we focused on a simple comparison task. We would highlight two words within a visualization containing words of different sizes and ask subjects to choose the one with the larger font size. While value-level accuracy in judging font size seems unnecessary for many high-level interpretations, the ability to make accurate *relative* judgements of represented data is important. Unless users can reliably discern that words with higher values are bigger than

those with lower values, the relationships between data associated with these words will be distorted or lost. We believe that decently accurate perception of relative size is a prerequisite even for such high-level tasks as gist-forming and summary comparison, in addition to the more obvious ones of order reading and value reading. Therefore, though users in the wild are rarely faced with a single pairwise comparison, we believed performance at this task would help us measure the ability to perform higher level tasks that rely on the same perceptual abilities.

There were other tasks that we considered, as well. One solution might have been to ask participants to make an absolute judgment of font size (e.g., 1.5mm), or to compare to a memorized baseline size (e.g., bigger than baseline). Although such tasks are simple, their detachment from the context of real-world tasks might have lead to idiosyncratic strategies, such as focusing attention on the height of a single letter instead making a holistic judgement about a whole word. At the other extreme, another solution might have been to ask which word in an entire cloud has the biggest font, while systematically manipulating the distribution of font sizes within that cloud. However, this task presents many degrees of freedom that make precise measurement more difficult. For example, it is not clear whether we should measure precision as the difference between the biggest font versus the next biggest, of versus the algebraic or geometric mean of the distribution, or versus some other property of the distribution [17], [18], [19]. We chose to use the pairwise comparison task in most of our experiments for the greater control it offered us. After having explored perceptual biases in this task, however, we still wanted to be sure that what we had found was extensible to more real-world situations, and so we ran a set of experiments using the pick-the-biggest-word task, which showed similar results (see Section 7).

4 General Experimental Design

As discussed in Section 3, we focused on *comparative* judgements of size rather than exact ones. In particular, we focused on the use of word clouds. Not only are these one of the most common mediums for font size encodings, but they also present a challenging context for reading values, given the dense proximity of distracting words and the frequent lack of alignment to any shared baseline for any pair of words.

4.1 Task Setup and Measures

Participants were first given instructions on the task, and read a tutorial indicating the difference between a word's font size and the area it took up on the screen. Participants were instructed to complete the tasks as accurately as possible.

Across multiple experiments, we gave participants the following task with different stimuli: Upon being shown a word cloud in which two words were highlighted using a darker gray, participants were asked to click on the highlighted word that had been given the larger font size. We were sure to fully explain the distinction between font size and the general footprint of a word on the screen. While others have observed instances of users misinterpreting the *meaning* of font size encodings [3], we were concerned primarily with perceptual abilities, and so did not want there to be any confusion for participants.

For each task, we recorded which word the participant clicked, as well as the time it took. We measured time only to test for fatigue effects (were tasks getting slower over time, or was performance decreasing)—our primary measure was accuracy. We used analyses of variance (ANOVAs) to test for differences among participant accuracies across conditions. Upon clicking a word, the participant was immediately presented with the next trial.

4.2 Factor Agreement

In each experiment, we tested a potentially biasing word factor to see if it affected the perception of font size. These factors were features of the words that vary based on the *contents of the words themselves*, such as word length, rather than attributes of the font that could feasibly be controlled across the entire visualization. To check for bias of a factor, we employed a method we have called **factor agreement**.

Factor agreement indicates whether the difference in the factor in question *reinforces* or *opposes* the difference in font size (see Figure 2). For example, if the word within a given pair with the larger font size also contains more letters, then we would say that word length **agrees** with font size. However, if the word with the larger font size contains fewer letters, we would say word length **disagrees** with font size. If both words are the same length, then the word length factor is **neutral**. It is not necessarily the case that any given factor's agreement or disagreement will affect a user's perception of font size, but if they do have an effect, we would expect user accuracy to decrease in situations of disagreement.

4.3 Stimuli

Stimuli for these experiments were all generated within a web browser. For early experiments, we created our own clouds using the D3 visualization library [20]. In later experiments, to create more realistic scenarios, we used jQCloud [21], a word cloud library that packs words more densely using a spiral layout. With the exception of Experiment HEIGHT3, in which we explicitly decided to test a sans serif font (see Table 1), we used Times New Roman for all of our stimuli.

The words used in each experiment were either English words or "pseudowords" (see Table 1). Pseudowords were constrained strings of random characters we created for greater control over the character glyphs being used and to factor out any semantic weight. Precise characteristics of these pseudowords varied between experiments (see Section 5). When building word clouds with English words, we drew from the Corpus of Contemporary American English (COCA) [22]. We built a database that allowed us to query for words with specific attributes (e.g., length).

The two **target words** between which participants had to choose varied in their font sizes and attributes from experiment to experiment. They were also joined by 40 **distractor words** in each stimulus, whose sizes were distributed across a normal distribution. After some calibration through pilot studies, we kept the difference in font size between the two target words relatively small. Accuracy was high enough in these conditions that testing larger differences was deemed unnecessary.

One issue that came up during experimentation was how different browsers perform subpixel-rendering. For non-integer font sizes (e.g., 12.5px), modern browsers sometimes use different rendering methods that can result in participants with different machines viewing slightly different sizes. However, as a between-subjects factor, browser differences should not affect the within-subjects factors that make up most of the factors in our experiments. Additionally, the experiments we chose to report in the main body of the paper all used integer-value font sizes. However, it is worth noting that some of the between-subjects

	Factor agreement					
Factor	agree		neutral		disagree	
word length	hello sam	bigger font, longer word	hello world	same length	hello goodbye	bigger font, shorter word
word height	help corn	bigger font, taller word	plot flop	same "raw height"	corn help	bigger font, shorter word
word width	joyful letter	bigger font, wider word	litter fillet	same "raw width"	little hummed	bigger font, narrower word

Fig. 2. In this figure, we show examples of the different conditions of **factor agreement** (see §4.2) for the three main factors of word shape that we tested: word length, word height, and word width. For height, we were concerned with the use of tall and short characters, rather than height differences resulting from font size. Similarly, for word width, our primary concern was not the *final* width of the word in the stimulus, but rather the *raw width*—its width before any changes in font size had been applied. While "litter" is wider than "fillet" in the above figure, they are the same width when written in the same font size.

Fig. 3. For many of our experiments, we used word clouds that we built using the D3 visualization library [20]. These clouds dispersed words randomly throughout the two-dimensional space, restricted only by avoiding collisions with the borders and other words. Words were either drawn from the English words within COCA [22] or pseudowords created using random characters (as shown here).

effects described in the supplemental materials may be influenced by cross-browser differences.

4.4 Participants

Over 12 experiments, we recruited 301 participants using Amazon's Mechanical Turk framework, restricted to native English speakers residing in North America with at least a 95% approval rating. These participants ranged in age from 18 to 65 (with a mean of 33) and were made up of 172 males and 129 females. We paid participants either $1.00 or $2.00 for their time, depending on the number of stimuli with which we presented them (which varied from 56 to 150).

It is worth noting that by using a crowdsourced framework, we sacrifice control over a number of environmental factors that could affect a participant's perception. These include browser differences (as discussed above), along with things like viewing distance, lighting, etc. Such factors may have influenced differences between participants, and may be worth investigating in future in-person studies. However, we believe we can rely on them being relatively consistent for individual participants, and therefore they should not affect the reported within-subjects factors.

To account for the varying levels of engagement often seen in participation in online studies, we followed acknowledged best practices to improve the reliability of our experimental data, including randomizing questions and manually checking for "click-through" behavior [23], [24]. Within each session, we also included "validation stimuli" with font size differences of a full 10 pixels. These validation stimuli were used as engagement checks to verify that participants had properly understood the instructions and were giving earnest effort. These questions were not considered in further analysis.

5 Exploring Biasing Factors

Over the course of our explorations, we ran over a dozen experiments involving hundreds of participants on Amazon's Mechanical Turk. Rather than describe the results for every experiment in detail, we have organized the main results and takeaways from each experiment into Tables 1 and 2 and will discuss a subset of them in greater depth in this section. The remaining experiments are described in full in the supplemental materials. We have structured the experiments by the main factors that we tested for bias: word length, character height, and word width (shown in Figure 4).

Label	E/P	Effect of Δ font size	Primary bias factor	Effect of bias factor agreement	Additional factor	Accuracy at min Δ font size			Notes
						agree	neutral	disagree	
len1	P	✓	word length†	✓	-	0.860	0.879	0.753	Word length biases perception of font size
len2	P	✓	word length†	✓	base font size‡	0.861	0.816	0.734	We see a greater bias at larger base font (30px vs. 20px)
len3	P	✓	word length†	✓	base font size†	0.825	0.838	0.642	Tested wider variety of baseline font sizes
len4	E	✓	word length†	✓	-	0.992	0.942	0.867	Bias still present with English words and denser word clouds
height1	P	✓	word height†	✓	-	0.974	0.909	0.684	Character heights bias perception of font size
height2	P	✓	word height†	✓	-	0.929	0.810	0.529	Proportional difference in font size seems to matter more than absolute difference
height3	P	✓	word height†	✓	-	0.937	0.795	0.525	Bias still present when word clouds use sans serif font
height4	P	✓	word height†	✓	base font size†	0.931	0.790	0.479	We see a greater bias at larger base font (30px vs. 20px)
height5	P	✓	word height†	✓	base font size‡	0.963	0.854	0.489	Accuracy hits ceiling between 20-25% size difference
width1	E	✓	word width†	✓	-	0.975	-	0.909	Bias present when length is held constant and width varies
width2	E	✗	word length†	✗	-	0.982	-	0.982	No bias when width is held constant and length varies
box1	E	✓	word width†	✗	-	0.914	0.932	0.908	No bias with corrected-width rectangular bounding boxes
big1	P	✓	word length†	✓	number of near misses	0.888	0.826	0.658	Tested using "pick the biggest word" task
big2	P	✓	word length†	✓	number of near misses	0.811	-	0.562	Tested wider variety of length differences
† - within-subjects factor ‡ - between-subjects factor									

TABLE 1
An overview of the experiments we ran for this study. Each experiment compared at least two factors: the difference in font size between the two target words, and a potentially biasing factor that was a feature of the words' shape. (Additional factors tested are described in the supplemental materials.) Here, we report the effects of these factors and the effect size of factor agreement at the smallest difference in font size tested (generally a 5% difference). Experiments with a white background are described in Sections 5 and 6, while those with a gray background are described in full in the supplemental materials. In column "E/P", "E" indicates that English words were used and "P" indicates that "pseudowords" were used (see §4.3).

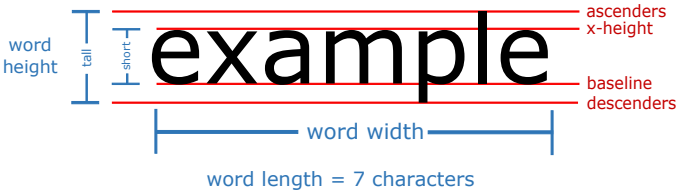


Fig. 4. We looked for biasing effects on font size perception for three main factors of word shape (shown here in blue): word length (§5.1), word height (§5.2), and word width (§5.3). For our experiments on height, words were broken down into two categories: "tall" words containing both ascenders and descenders, and "short" words whose height was contained between the font's baseline and x-height.

5.1 Word Length

The first attribute we tested was **word length**: the number of characters contained within a word. Longer words take up more space, and have a larger *area* than shorter words of the same font size, and even some shorter words with *larger* font sizes. We predicted that these differences in area could interfere with the ability to perceptually distinguish words by pure font size alone.

We ran four total experiments using word length as a test factor. In each one, we observed a significant effect in which participant accuracy went down when word length disagreed with font size. The details for these experiments can be found in Tables 1 and 2, as well as the supplemental materials. We will describe two of the most important experiments here.

5.1.1 *Experiment LEN1*
Word length biases perception of font size

For our first experiment on word length, we presented participants with word clouds of our own creation as described in Section 4.3 (see Figure 3). To afford greater control in stimulus generation, we used words of random characters, excluding characters with ascenders or descenders (e.g., "h" or "g"—see Figure 4) as well as characters of abnormal width (e.g., "w" or "i"). We enforced a minimum distance between the two highlighted words, and ensured that they shared no common horizontal or vertical baselines that would aid in comparison.

We tested two main factors: font size and word length. Both were examined using within-subject comparisons. Font size for the first target word was either 20px, 21px, or 22px, while font size for the second word was either 20px or 22px. Length for both target words alternated between 5 characters and 8 characters. The full combination of these factors created 24 conditions, of which 16 had a "correct answer" (i.e., one of the words had a larger font size), and 8 of which did not (i.e., the words were the same font size). This allowed us to observe both instances of factor agreement and disagreement, as well as see which way people leaned at the extreme marginal case where the sizes were equal.

We tested 31 participants, each of whom saw 150 stimuli (6 per each of the 24 conditions described above, as well as 6 engagement

Experiment	N	Factors	Conditions	Analysis of Variance				
				W/B	df1	df2	F	p-value
len1	31	Δ font size	w1: [20, 21, 22px], w2: [20, 22px]	W	1	150	59.21	< 0.0001
		word length agreement	w1: [5, 8 chars], w2: [5, 8 chars]	W	2	150	14.91	< 0.0001
len2	39	Δ font size	[5, 10, 15, 20%]	W	3	418	58.96	< 0.0001
		word length agreement	w1: [4, 7, 10 chars], w2: [4, 7, 10 chars]	W	2	418	12.13	< 0.0001
		base font size	[20, 30px]	B	1	37	7.98	0.008
len3	20	Δ font size	[5, 10, 15, 20%]	W	3	926	85.43	< 0.0001
		word length agreement	w1: [5, 8 chars], w2: [5, 8 chars]	W	2	926	31.60	< 0.0001
		base font size	[20, 25, 30, 35px]	W	3	926	8.57	< 0.0001
len4	20	Δ font size	w1: [20px], w2: [21, 22, 23, 24px]	W	3	269	7.84	< 0.0001
		word length agreement	w1: [5, 8 chars], w2: [5, 8 chars]	W	2	269	14.32	< 0.0001
height1	32	Δ font size	w1: [20, 21, 22px], w2: [20, 22px]	W	1	155	55.31	< 0.0001
		word height agreement	w1: [tall, short], w2: [tall, short]	W	2	155	71.22	< 0.0001
height2	20	Δ font size	w1: [20, 22, 24px], w2: [21, 23px]	W	5	323	45.88	< 0.0001
		word height agreement	w1: [tall, short], w2: [tall, short]	W	2	323	83.90	< 0.0001
height3	20	Δ font size	w1: [20, 22, 24px], w2: [21, 23px]	W	5	323	59.42	< 0.0001
		word height agreement	w1: [tall, short], w2: [tall, short]	W	2	323	36.10	< 0.0001
height4	20	Δ font size	[5, 10, 15, 20%]	W	3	448	59.81	< 0.0001
		word height agreement	w1: [tall, short], w2: [tall, short]	W	2	448	88.39	< 0.0001
		base font size	[20, 30px]	W	1	448	44.9	< 0.0001
height5	40	Δ font size	[5, 10, 15, 20, 25%]	W	4	546	94.39	< 0.0001
		word height agreement	w1: [tall, short], w2: [tall, short]	W	2	546	207.2	< 0.0001
		base font size	[20, 30px]	B	1	38	20.09	< 0.0001
width1	20	Δ font size	w1: [20px], w2: [21, 22, 23, 24px]	W	3	133	6.77	0.0003
		word width agreement	[+10px, -10px]	W	1	133	11.33	0.001
width2	19	Δ font size	w1: [20px], w2: [21, 22, 23, 24px]	W	3	126	1.47	0.23
		word length agreement	[+3 chars, -3 chars]	W	1	126	0.00	1.00
box1	20	Δ font size	[5, 10, 15, 20%]	W	3	209	10.88	< 0.0001
		word width agreement	[-20px, 0px, +20px]	W	2	209	0.52	0.60
big1	19	Δ font size	[5, 10, 15, 20%]	W	3	414	5.82	0.0007
		word length agreement	target: [5, 8 chars], near misses: [5, 8 chars]	W	2	414	10.10	< 0.0001
		# near misses	[1, 4]	W	1	414	33.66	< 0.0001
big2	19	Δ font size	[5, 10, 15, 20%]	W	3	846	3.02	0.03
		word length agreement	[-5, -3, -1, 1, 3, 5 chars]	W	5	846	8.00	< 0.0001
		# near misses	[1, 4]	W	1	846	7.00	0.008

TABLE 2
An overview of the statistical tests we ran for this study. For each experiment, we show the number of participants (N), the factors and their levels (specifying conditions for both target words—w1 and w2—where appropriate), whether the factors were treated as within- or between-subjects factors, and the analyses of variance for each. Effect sizes can be seen in Table 1 and in the supplemental materials.

tests). While this initially seemed like a large number of stimuli, we saw no fatigue effects in any of our studies. Average time to completion was 5.8 minutes, and the comments we received from participants were positive. We analyzed answers to questions with a correct answer and without a correct answer separately.

For data where there was a correct answer, we calculated the font size difference (1 or 2 px) and word length agreement ("agree," "neutral," or "disagree") for each stimulus. We then ran a two-way analysis of variance (ANOVA) to test for the effect of the font size difference and word length agreement. We saw main effects for both font size difference ($F(1,150) = 59.21$, $p < 0.0001$) and word length agreement ($F(2,150) = 14.91$, $p < 0.0001$). Specifically, participant performance decreased when the difference in word length *disagreed* with the difference in font size, as well as when the difference in font size was smaller (see Figure 5). A post hoc test using Tukey's HSD showed that the "disagree" condition was significantly different from both the "neutral" and "agree" condition, though the latter two were not statistically distinguishable from one another.

For data where there was no correct answer, we tested to see if the rate at which participants picked the *longer* of the two words was significantly different from chance. Specifically, we calculated the rate at which each participant picked the longer of the two words when the font sizes were the same ($M = 0.59$, $SD = 0.17$) and ran a two-tailed, paired Student's t-test to compare these

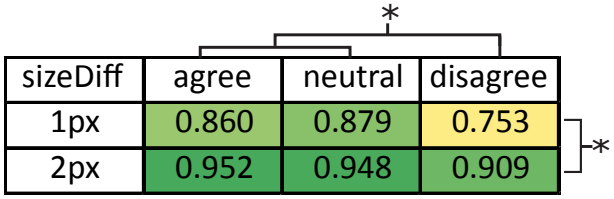


Fig. 5. This table shows the average participant accuracy for each combination of factors for experiment LEN1 (§5.1.1). A two-way ANOVA showed significant main effects for both size difference and length agreement. A post hoc Tukey's HSD test showed that the "disagree" condition (i.e., when the longer of the two words had the smaller font size) was significantly different from the "agree" and "neutral" cases, though the latter two were not distinguishable from one another.

values against an equally sized collection of values of 50%. We found that participants were significantly more likely to pick the longer of the two words ($t(30) = 2.99$, $p = 0.005$), indicating the same direction of bias as seen with the data with correct answers.

5.1.2 Experiment LEN4
Biases still present with full English words

For this experiment, we wanted to test whether the effects that we had seen using "fake" words and our relatively sparse word clouds would still be present in a more realistic setting. Specifically, rather than generating random strings of characters for words, we used

sizeDiff	agree	neutral	disagree
5%	0.992	0.942	0.867
10%	1.000	1.000	0.917
15%	0.992	0.992	0.992
20%	0.992	1.000	0.975

Fig. 6. This table shows the average participant accuracy for each combination of factors for experiment LEN4 (§5.1.2), in which we looked for a bias of length agreement within a more realistic collection of word clouds. After a two-way ANOVA showed significant main effects for both length agreement and font size difference, post hoc tests showed that the "disagree" condition and the closest font size difference were the real departures from the rest of the conditions.

Fig. 7. To create a more realistic context for experiment LEN4 (see §5.1.2), we used a modified version of the jQCloud library to create stimuli [21]. These word clouds were more densely packed, more closely resembling what participants might be used to seeing in other settings.

words drawn from the COCA [22]. We also switched from our own word cloud implementation (Figure 3) to a modified version of a commonly used library called jQCloud [21] (Figure 7). These clouds packed words more densely by using the spiral positioning layout. The jQCloud library also allowed us to easily modify the aesthetics of the clouds through CSS, creating images more closely resembling the types of word clouds participants might be familiar with seeing in other contexts, such as Wordles [3].

Our factors were once again font size and word length, each a within-subject factor by our design. We held the first target word at a font size of 20px while the second word's font size was either 21px, 22px, 23px, or 24px. The word length of each target word alternated between 5 and 8 characters. All words were restricted to characters that contained no ascenders or descenders to avoid any effects resulting from height. The full combination of these factor levels resulted in 16 combinations—each, in this case, with an explicitly correct choice.

We tested 20 participants, each of whom saw 102 stimuli (6 per each of the 16 conditions, plus an additional 6 engagement tests). After calculating the font size difference and word length agreement for each stimulus, we ran a two-way ANOVA to test for the effect of these two metrics. Once again, we saw main effects for both font size difference ($F(3,269) = 7.84$, $p < 0.0001$) and word length agreement ($F(2,269) = 14.32$, $p < 0.0001$), indicating lower accuracy in instances of word length disagreement at close font sizes (see Figure 6). Post hoc tests with Tukey's HSD identify the "disagree" condition and the closest font size difference as the main departures from the rest of the conditions. The lack of difference between the higher-scoring conditions may be the result of ceiling effects, as accuracy was very high across the board.

5.1.3 Discussion

In these experiments, we see a very consistent bias towards longer words. Word length, it appears, does affect user perception of font size. However, accuracies across both experiments were higher than we had been anticipating. With mean accuracies consistently near or above 90%, participants seemed surprisingly good at making these comparisons. These high accuracies may have created a ceiling effect, which could account for the lack of distinction between the "agree" and "neutral" conditions in post hoc tests. Dips in accuracy, while consistent, happened primarily at very close font sizes, but even then participants did notably better than chance. This may be cause to *trust* user perceptions of font size encodings. However, the number of letters is just one of many features that factors into the diversity of shapes words can make.

5.2 Word Height

The next potentially biasing feature of a word that we tested was a word's *height*. Specifically, there are some characters in the basic Latin alphabet that are taller than others due to the presence of **ascenders** and **descenders** in their glyphs. Ascenders—found for example in the letters "h" and "k"—are marks that reach above a font's *x-height*, while descenders—as in "g" and "y"—extend below a font's baseline (see Figure 4). Given that height is perhaps the easiest way to tell font sizes apart when comparing words of varying lengths, we wanted to see whether the presence or lack of such characters would adversely affect user judgement.

We ran five experiments investigating this possibility, and saw a significant bias for character height in each of them (see Table 1). We will again discuss the most important of these experiments here and relegate the others to the supplemental materials.

5.2.1 Experiment HEIGHT1
Character heights bias perception of font size

For our first experiment investigating the effect of character height, we again used words of random characters to give us fine-tuned control over the characters present. We defined two types of "fake" words: **tall** and **short**. Short words were generated using only characters without ascenders and descenders (e.g., "a" or "c") and excluding characters of abnormal width (e.g., "w" or "i"). For tall words, we used the vowels "a", "e", "o" and "u" and added characters with ascenders and descenders, again excluding tall characters with abnormal width (e.g., "f", "j", "l"). Short words are naturally rectangular since all of their characters share the same height, but the ascenders and descenders in tall words unbalance this rectangular shape. In order to balance the tall words' shapes, we positioned tall characters both in the beginning and end of the word making sure that if a word started with an ascender, it would end with a descender and vice-versa. Each tall word was made up of 8 characters: 3 short characters and 5 tall characters.

We used precisely the same experimental setup as in Section 5.1.1, with the factor of word length exchanged for word height: the presence or absence of ascending and descending characters.

sizeDiff	agree	neutral	disagree
1px	0.974	0.909	0.684
2px	1.000	0.965	0.932

Fig. 8. This table shows the average participant accuracy for each combination of experimental factors for experiment HEIGHT1 (§5.2.1). A two-way ANOVA showed main effects for both word height agreement and font size difference. Post hoc analysis using Tukey's HSD showed that all experimental conditions were statistically distinguishable from one another. Most notably, accuracy is lowest for the "disagree" condition with the closest difference in font size.

This meant that the first target word again varied between sizes of 20px, 21px, and 22px while the second word varied between 20px and 22px as both words alternated back and forth between the tall and short words. Of the 24 conditions created by combining these factors, 16 had a difference of font size (and therefore a "correct" answer) while 8 did not. We analyzed the data for stimuli with a correct answer and stimuli without one separately.

For data where there was a correct answer, we calculated the font size difference (1 or 2 px) and word height agreement ("agree," "neutral," or "disagree") for each stimulus. We then ran a two-way ANOVA to look for effects of these metrics on participant accuracy. We saw significant main effects for both height agreement ($F(2,155) = 71.22$, $p < 0.0001$) and font size difference ($F(1,155) = 55.31$, $p < 0.0001$). These effects went in the same direction as seen in Section 5.1 with word length: accuracy dropped when character height *disagreed* with font size and when the font sizes were particularly close (see Figure 8). Post hoc tests with Tukey's HSD showed all pairwise combinations of conditions to be statistically significant.

For data without a correct answer, we calculated the rate at which each participant picked the *tall* word when presented with two words of the same font size ($M = 0.67$, $SD = 0.07$) and compared these values to a collection of 50% values with a two-tailed, paired Student's t-test. We saw that participants chose the taller of the two words at a significantly higher rate than chance ($t(31) = 12.91$, $p < 0.0001$).

5.2.2 Discussion

Like word length, character height seems to create a consistent bias on participant perception of font size. In fact, the bias for character height seems to be more pronounced, with accuracy in the worst cases dropping to levels not much better than chance (see Table 1). However, instances of these height differences are relatively rare in English. The list of words we used from COCA [22] has in total 25,859 eligible words after removing duplicates and words containing numerals and punctuation. Of these, only 870 fit our definition of "short" words—approximately 3.3% of eligible words. As such, the extreme comparison of tall to short words would likely not happen often in the wild. However, there are less extreme comparisons—words containing only a few ascenders or descenders, words containing only one or the other, etc.—that may be more common and still exhibit this bias.

5.3 Word Width

After running our tests on word height, we decided to look for the the effect of a different factor: word *width*. In our height experiments, we held length constant and attempted to control for width by excluding characters of abnormally small or large width (as described in Section 5.2.1). However, there were still small differences in glyph widths even outside of those characters, which created variance in width from word to word, even within the same length conditions. In a post hoc test, we computed a **width agreement** metric for each stimulus from experiment HEIGHT2 indicating whether the difference in width went in the same direction as the difference in font size. It was only for stimuli with the smallest font size difference that we saw any width disagreement, given that we had attempted to make widths neutral. We ran a two-way ANOVA looking for an effect of width agreement, specifically on the stimuli in the closest font difference case. The effect we saw was significant ($F(2,38) = 13.73$,$p < 0.0001$). Accuracy in the disagree condition ($M = 0.523$, $SD = 0.18$) was substantially lower than accuracy in the agree condition ($M = 0.82$, $SD = 0.10$).

This led us to an interesting question. We knew that longer words created a bias for font size perception, as described in Section 5.1, but we did not know *why*. Was this bias the result of longer words taking up more space, and therefore a function of width, or were participants actually making a numerosity judgement about the letters? We hypothesized that the main factor in this effect was width rather than length, thinking that words—especially *real* ones—are read more or less as a whole, rather than letter by letter [25]. To test this hypothesis, we ran two additional experiments to isolate the effects of width and length.

5.3.1 Experiment WIDTH1
Bias present when width varies but not length

In our first of these experiments, we wanted to see whether word width biased font size perception even when the number of characters and character height were held constant. Varying width but not length put a tight constraint upon the words we were able to use; differences between character widths are small, and so words that differ substantially in one factor but not the other are rare. For our stimuli, we chose a collection of pairs of words that were each 8 characters long, but differed in **raw width** by 10 pixels. We defined "raw width" to be a word's width computed at a font size of 20px, so that we could have a measure of width differences that was separate from our font size factor. We also made sure that each pair of words shared the same character height.

Our two factors for this experiment were width agreement and font size difference. For each stimulus, one of the target words had a font size of 20px, while the other was either 21px, 22px, 23px, or 24px. For the width agreement factor, the larger of the two words either had a raw width that was 10 pixels greater than the smaller word ("agree") or 10 pixels less than the smaller word ("disagree"). Four font size differences combined with two levels of width agreement gave us 8 conditions, each of which had a "correct" answer.

We tested 20 participants, each of whom saw 56 stimuli (6 per each of the 8 conditions, as well as 6 engagement tests). After calculating the font size difference and width agreement of each stimulus, we ran a two-way ANOVA to test for the effects of the two factors on participant accuracy. We saw main effects for both width agreement ($F(1,133) = 11.33$, $p = 0.001$) and font size difference ($F(3,133) = 6.77$, $p = 0.0003$) indicating a drop off in accuracy for width disagreement at close font sizes (see Figure 9). While a post hoc Tukey's HSD test only showed the smallest size difference condition to be statistically distinguishable, this

sizeDiff	agree	disagree
5%	0.975	0.909
10%	1.000	0.992
15%	0.992	0.992
20%	1.000	0.983

Fig. 9. This table shows the average participant accuracy for each combination of experimental factors for experiment WIDTH1 (§5.3.1). In this experiment, target words had a difference of 10 pixels in raw width (i.e., their width at the same font size). In the "agree" condition, this width difference was in the same direction as the difference in font size, while it was in the opposite direction for the "disagree" condition. A two-way ANOVA showed significant main effects for both width agreement and font size difference. Only the lowest size difference was statistically distinguishable in post hoc tests, perhaps due to ceiling effects given the very high overall accuracy.

may have been due to ceiling effects, given the very high accuracy across all other conditions.

5.3.2 Experiment WIDTH2
*Bias **not** present when length varies but not width*

In the second of these experiments, we wanted to see whether the number of letters in a word had any effect on font size perception outside of the correlated factor of width difference. For our stimuli, we chose pairs of words that had the same raw width (described in Section 5.3.1) but differed by 3 letters in length. Of the words we had available from which to choose, this was the largest length difference that provided us with enough pairs. Each pair of words shared the same character height, as well.

Our two factors for this experiment were length agreement and font size difference. Once again, one of the two target words in each stimulus had a font size of 20px, while the other was either 21px, 22px, 23px, or 24px. For the length agreement factor, the larger of the two words had either 3 more characters than the smaller word ("agree") or 3 fewer characters than the smaller word ("disagree"). Four font size differences combined with two levels of length agreement gave us 8 conditions, each of which had a "correct" answer.

We tested 19 participants, each of whom again saw 56 stimuli. After computing the font size difference and length agreement of each stimulus, we ran a two-way ANOVA to test for the effects of these factors on participant accuracy. This time, we saw no main effects for either font size difference ($F(3,126) = 1.47$, $p = 0.23$) or length agreement ($F(1,126) = 0.00$, $p = 1.00$). Accuracy was quite high across all conditions (see Figure 10). This seems to indicate that any bias created by number of letters alone is not strong enough to register without also varying the stronger factor of word width.

5.3.3 Discussion

The restriction of varying only *one* of width and length meant that we were not able to test very large differences in either factor. As such, we did not expect to see a vary large effect size for either experiment. However, from these results, we feel we can conclude that width is the more important factor to consider when worrying about bias. Length may matter in some extreme cases, but we stretched the degree to which length can vary without width to the limits of the English language, and still saw no effect. Practically, therefore, width seems the more relevant concern.

sizeDiff	agree	disagree
5%	0.982	0.982
10%	1.000	0.991
15%	0.991	1.000
20%	0.982	1.000

Fig. 10. This table shows the average participant accuracy for each combination of experimental factors for experiment WIDTH2 (§5.3.2). In this experiment, target words had a difference of 3 characters in their length (going with or against the direction of the difference in font size in the "agree" and "disagree" conditions, respectively). A two-way ANOVA showed no significant main effects for either factor, and accuracy was very high across the board.

6 Debiasing with Rectangles

In Section 5, we show that there are multiple ways in which a word's shape can bias interpretation of its font size. Depending on the task a designer intends a user to undertake, the effect of this bias may not be large enough to warrant much intervention—a possibility we discuss further in Section 8. However, for tasks precise enough to be concerned by these effects, the next question is what we can do as designers to *mitigate* this bias.

One potential method for this debiasing effort was inspired by the work of Correll et al. debiasing area discrepancies in tagged text [26]. In this work, the authors determined that users suffered from an area bias when making numerosity judgements of words tagged with colored backgrounds. Specifically, when the number of words disagreed with the *area* of the colored backgrounds, accuracy dropped dramatically. However, they were able to counteract this bias by adjusting the area of the backgrounds for underrepresented words.

We suspected that such a technique could be useful for the biases we observed in font size encodings. By enclosing individual words in filled bounding boxes, we can create a redundant encoding for font size that may alleviate the issue of diverse word shapes. These bounding boxes would also give us a glyph whose proportions we can adjust without fearing any change in legibility.

As such, we decided upon the following potential debiasing technique: We would surround each word with a **padded bounding box**. These boxes would contain the full height of any potential character, going from the ascender line to the descender line (see Figure 4). The width of each box would be adjusted such that they all shared the same raw width—which is to say, they would be equal in width if they all contained words of the same font size. With such padding, the difference in rectangle width and height would always agree with the font size difference for any two words, creating a more reliable and readable indication than the word alone. We ran an experiment to test whether this strategy would help increase user accuracy in cases of factor disagreement.

6.1 Experiment BOX1
Can debias encoding with rectangular highlights

To test our debiasing technique, we ran an experiment with a similar design to that described in experiment LEN4 (described in Section 5.1.2). The factors for our stimuli were font size difference (which varied in increments of 5, 10, 15, and 20% from a base font of 20px) and word length (which alternated between 5 and 8 characters for each word). For this experiment, we also ensured that whenever the two target words were the same length, they also

Fig. 11. By containing each word in a color-filled bounding box and padding the sides of each bounding box such that their widths were proportional to their font sizes, we were able to eliminate the effect of width disagreement.

sizeDiff	agree	neutral	disagree
5%	0.914	0.932	0.908
10%	0.983	0.992	0.933
15%	0.983	0.971	0.992
20%	0.992	0.996	0.983

*

Fig. 12. This table shows the average participant accuracy for each combination of experimental factors for experiment BOX1 (§6.1). In this experiment, words were given padded bounding boxes (as in Figure 11) in an attempt to mitigate the bias created by disagreement in word width. While a two-way ANOVA showed there to be a significant main effect of size difference on accuracy, no main effect was seen on word width agreement—indicating that padded bounding boxes may be a viable way of debiasing font size perception.

had the same raw width, and when they were not the same length, they had a difference in raw width of 20 pixels. These factor levels created 16 conditions, each of which had a "correct" answer.

Rather than showing participants a pure word cloud, we placed padded bounding boxes around each word (see Figure 11). These bounding boxes were padded on either side such that the rectangle for each word had the same raw width before any differences in font size had been applied. Participants were instructed in the tutorial that the rectangles containing the words were sized proportionally to the words' font sizes.

We tested 20 participants, each of whom saw 102 stimuli (6 for each of the 16 conditions, plus an additional 6 engagement checks). After computing the length/width agreement and font size difference of each stimuli, we ran a two-way ANOVA to test for the effects of these factors on participant accuracy. While we found a significant main effect for font size difference as before ($F(3,209) = 10.88$, $p < 0.001$), we saw no effect of length/width agreement ($F(2,209) = 0.52$, $p = 0.60$). Even in the typical worst case—conditions with factor disagreement and the smallest difference in font size—participants scored over 90% accuracy (see Figure 12). To this degree, it seems that the padded bounding boxes were successful at mitigating the bias introduced by length/width disagreement.

This technique of debiasing font size encodings is primarily a proof-of-concept. Aesthetically, word clouds like the one in Figure 11 are inferior to more standard layouts, and aesthetics can be an important factor to an encoding's utility [27]. It may be possible to create more aesthetic approaches, perhaps using other word features like font weight or tracking. At any rate, this shows that the effects of word shape on font size perception are possible to correct for.

Fig. 13. For experiments BIG1 (§7.1) and BIG2 (§7.2), participants were presented with word clouds of pseudowords and asked to pick the one with the biggest font size. In this example, "zoav" is the correct answer, with four near misses that are of longer length.

7 Alternate Task

A possible critique of this work is that our experimental task (pick the bigger of two highlighted words) does not necessarily reflect how font size encodings are used in the wild. Our reason for using this task was that it acts as a "visual primitive" for broader, more general tasks (see Section 3). It is not our intention to say that people routinely have to perform the act of comparing two words within a word cloud, but rather that the more high-level, interpretation-based tasks that people *do* perform rely upon this low-level perceptual ability.

Nonetheless, we wanted to confirm that the bias that we saw within the compare-two-words task was not specific to this precise experimental setup. In a further set of experiments, we looked for the same bias using a different task: finding the single biggest word within a cloud. While we believe that this task relies upon the same perceptual abilities as the comparison task, it is in some ways closer to how word clouds are used in practice. Picking out the biggest word (or words) from a visualization that uses font size to encode values is similar to the higher level task of asking what the data encoded by the visualization is "about."

To give us control over the gap in font size between target words similar to what we had in our previous experiments, we introduced a concept called **near misses**. Near misses are words that are *almost* as large as the biggest font size word, but not quite (see Figure 13). Explicitly controlling the near misses in each stimulus allowed us to evaluate multiple font size differences between the biggest word and the next biggest. It also gave us a new factor: the number of near misses.

Our general hypotheses for the pick-the-biggest task were that participant accuracy would be worse in instances of factor disagreement (as in our previous experiments), and that this effect

would be more pronounced in stimuli that contained *more* near misses to distract the participant.

7.1 Experiment BIG1
Bias still present in "pick the biggest" task

In our first experiment making use of the pick-the-biggest task, we sought to examine potential bias due to word length agreement or disagreement. We created a set of stimuli of word clouds made up of pseudowords (see Section 4.3). As before, stimuli contained 40 distractor words, in this case limited to font sizes below 40px. Stimuli then contained either 1 or 4 near miss words which were given a font size of 40px. Finally, each stimulus contained a target word (the "correct" choice) with a font size defined by a percentage increment above that of the near misses (either 5, 10, 15, or 20% bigger).

The factors for this experiment were font size difference (5, 10, 15, or 20%), target word length (5 or 8 letters), near miss word length (5 or 8 letters), and number of near misses (1 or 4). Each factor was varied within participants. The full combination of these factor levels resulted in 32 conditions. We tested 19 participants, once again recruited through Amazon Mechanical Turk, each of whom saw 134 stimuli (4 per each of the 32 conditions, plus an additional 6 engagement tests with a font size difference of 50%). After calculating font size difference and word length agreement for each stimulus, we ran a two-way ANOVA to test for the effect of the three metrics (including number of near misses). We saw main effects for all three factors: font size difference ($F(3,414) = 5.82$, $p = 0.0007$), length agreement ($F(2,414) = 10.10$, $p < 0.0001$), and number of near misses ($F(1,414) = 33.66$, $p < 0.0001$), indicating lower accuracy in instances of word length disagreement, more near misses, and closer font sizes (see Figure 14).

Our hypothesis that we would still see a biasing effect of length disagreement using a different task was confirmed. Interestingly, accuracies seemed to drop off even more when participants were performing the pick-the-biggest task than when they were performing the pairwise comparison task (see Figure 14). However, participants still achieved greater than 50% accuracy in each condition, performing better than chance.

7.2 Experiment BIG2
Wider variety of sizes in "pick the biggest" task

For a second experiment using the pick-the-biggest task, we were interested in whether the *magnitude* of the word length agreement or disagreement was relevant to the bias created—that is, would instances of greater disagreement hurt accuracy more than instances of small disagreement. We created a design that was similar to that described in Section 7.1, but with different levels for the word length disagreement factor. Rather than only considering words of 5 or 8 characters, we considered word length differences of 1, 3, and 5 characters in both the "agree" and "disagree" directions, for a total of 6 levels for this factor. We hypothesized that instances of large disagreement (e.g., 5 characters) would show lower accuracy than instances of small disagreement (e.g., 1 character).

We tested 19 participants on Amazon Mechanical Turk, each of whom saw 150 stimuli (3 per each of the 48 combinations of factors with an additional 6 engagement checks). We ran a two-way ANOVA to test for the effects of the three metrics, and again saw main effects for all three: font size difference ($F(3,846) = 3.02$, $p = 0.03$), length difference ($F(5,846) = 8.00$, $p < 0.0001$),

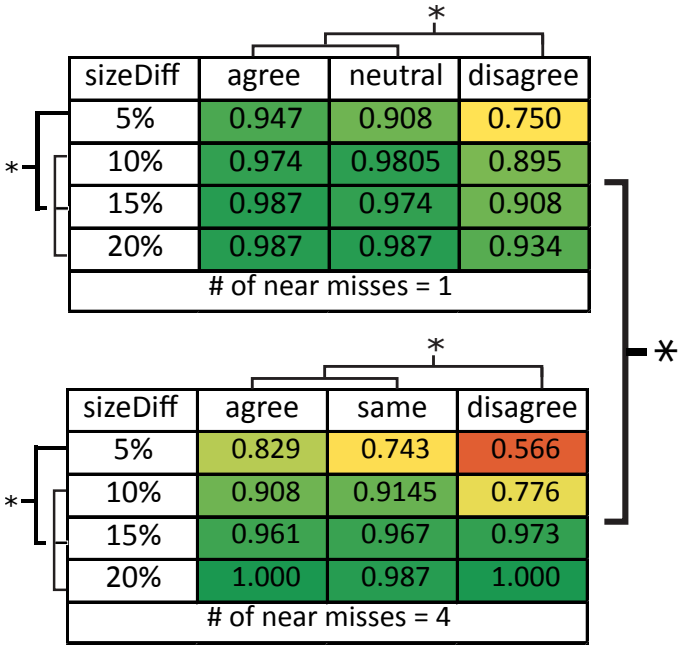


Fig. 14. This table shows the average participant accuracy for each combination of experimental factors for experiment BIG1 (§7.1). In this experiment, participants were asked to select the word with the largest font size. They were presented with word clouds containing a single word bigger than the rest (the "target" word) along with either 1 or 4 "near misses." A two-way ANOVA showed there to be a significant main effect for both the font size difference between the target and the near misses, for word length agreement, and for the number of near misses.

and number of near misses ($F(1,846) = 7.00$, $p < 0.008$)—each in the same direction as seen previously. We also noted, as expected, that accuracies were lowest in instances of largest disagreement and highest in instances of largest agreement (see Figure 15).

7.3 Discussion

The main takeaway from these two additional experiments is that the biasing effect of factor disagreement is not isolated specifically to the task of pairwise comparison, but can also be seen in a task that specifically tries to draw the user's attention to the most "important" word in the visualization. The detrimental effect of more "near misses" seems to perhaps indicate that while people are generally able to perform pairwise comparisons, needing to perform *multiple* of these can cause them to miss smaller words. However, performance is still better than chance in all but the most pathological cases.

8 Full Discussion

Results from other experiments not described above are laid out in the supplemental materials. In those experiments, we looked for a number of extra details and effects. We compared performance at different base font sizes. We tested to see if the results were the same with a sans serif font (which they were). We looked for a size difference ceiling past which participant accuracy maxed out (which proved to be between 20-25% size difference). Consistent across each experiment were the same things we saw in each of the experiments described in Sections 5, 6, and 7: decreased performance with factor disagreement at close size differences. It is worth noting that this effect is not simply the result of participants focusing on *area* rather than font size. Consider examples

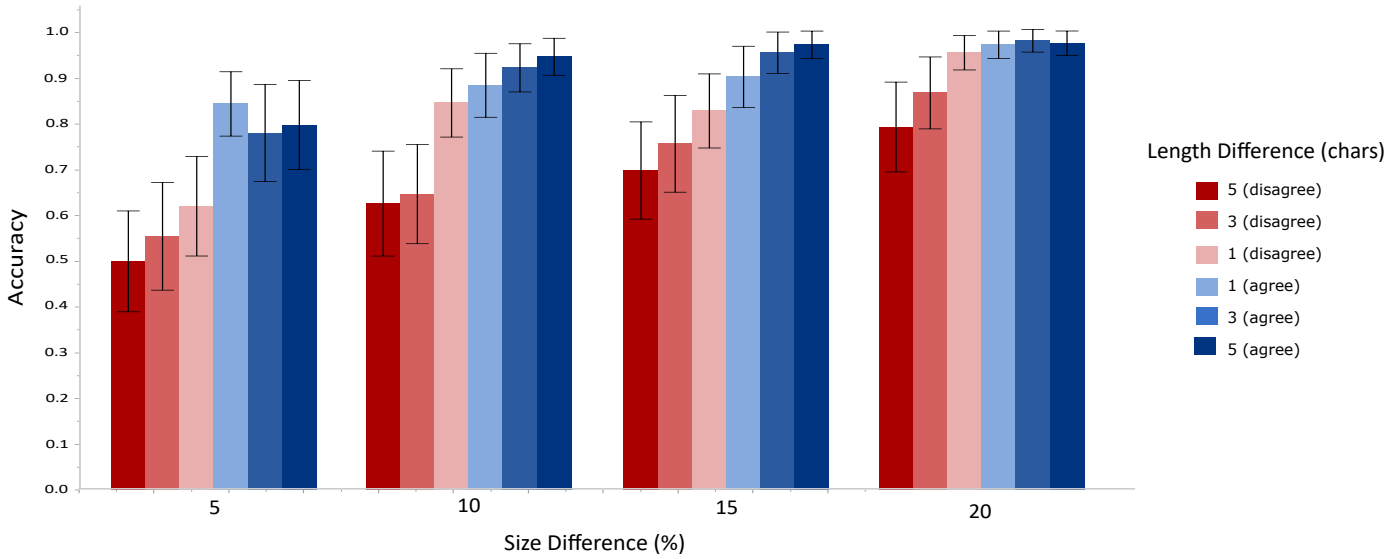


Fig. 15. This graph shows the average participant accuracy for combinations of experimental factors in experiment BIG2 (§7.2). In this experiment, participants were tasked with picking the word with the largest font size as in §7.1. We tested a wider variety of length differences, and saw that performance was generally lowest in cases of large disagreement and highest in cases of large agreement. These values are averaged across two levels of the "number of near misses" factor. Error bars represent a 95% confidence interval.

from our length disagreement experiments. While we observed decreased accuracy when a word with a 1-pixel-larger font size was significantly shorter than the other target, increasing the font size difference by a mere pixel resulted in very high accuracy—even though the difference in area disagreement created by this change in font size would be minimal.

Clearly, perceptions of font size can be biased by these factors. The relevant question for a designer is how much this bias will affect their end users, and whether it is worth designing around it. The effects that we saw occurred at very close differences in font size, and even then participants performed better than chance. It may be that our experimental setup artificially enhanced performance past what we would see in the wild—perhaps by having users focus in on two individual words out of many. Nonetheless, the consistently high accuracy that we saw across so many trials and conditions was remarkable. Despite the fact that font size encodings are rarely used for tasks requiring pixel-level accuracy, our findings seem to suggest that they may be more suitable for such tasks than previously thought. Given the particular utility of the font size encoding for textual data, expanding its potential uses could have significant impact. An important future direction of this work, therefore, will be to continue testing the limits of this perception in real-world applications.

While thorough investigation of these phenomena in more realistic contexts will be important for applying this work, it is also important to understand the psychophysical mechanism(s) responsible for the observed effects. Perceptual-level study of *why* this bias exists could help us predict whether effects might be better or worse in other viewing conditions, visualization contexts, or using different kinds of data. It may be useful for such future work to take the form of in-person studies for more precise measurement and better data gathering. This could also help validate our crowdsourced results in a more controlled environment.

Our debiasing attempts are a proof-of-concept, and show that it is possible to correct for the effects of factor disagreement in the event that a designer expects careful reading and comparison of their encodings. We believe there are more aesthetic ways of making these corrections, and are interested in exploring them further. Font weight, for instance, may interact with font size in ways that we could exploit in our encodings. Possible candidates for other methods include typeface modifications such as kerning, widths of individual letter glyphs, or even exploring the use of monospaced typeface (where all the characters have the same width causing words that have the same length to be the same width as well). Ultimately, whether or not debiasing is even necessary depends on how the encoding will be used in practice.

While we looked for biasing effects of a number of features related to a word's content—including length, width, character height, and font (see the supplemental materials)—there are more features that could be examined. These include color, font weight, and a word's semantic weight or meaning. Also, while we believe that the pairwise comparison and pick-the-biggest tasks allow us to get down to the perceptual primitives of higher level tasks, we are interested in testing a wider variety of tasks to better understand font size encodings in real world contexts.

9 Conclusion

We have explored the effects of different word shapes on the perception of data encoded through font size. Across multiple experiments, we have shown that the factors of word length, character height, and word width can all have a negative impact on one's ability to judge comparative font sizes, particularly when they differ in the opposite direction from the font sizes being compared ("disagreement"). These biases are consistent, but surprisingly small in their effects, possibly indicating that such encodings are better suited to higher accuracy tasks than previously expected. We have shown in a proof-of-concept design that correcting for them is possible by adjusting the visual encoding.

Acknowledgments

This work was supported in part by NSF awards IIS-1162037 and IIS-1162067, a grant from the Andrew W. Mellon Foundation, and funding from NSERC and the Canada Research Chairs program.

REFERENCES

[1] B. Y. Kuo, T. Hentrich, B. M. Good, and M. D. Wilkinson, "Tag clouds for summarizing web search results," in *Proceedings of the 16th international conference on World Wide Web*. ACM, 2007, pp. 1203–1204.

[2] C. Trattner, D. Helic, and M. Strohmaier, "Tag clouds," in *Encyclopedia of Social Network Analysis and Mining*. Springer, 2014, pp. 2103–2107.

[3] F. B. Viégas, M. Wattenberg, and J. Feinberg, "Participatory visualization with wordle," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 15, no. 6, pp. 1137–1144, 2009.

[4] S. Afzal, R. Maciejewski, Y. Jang, N. Elmqvist, and D. S. Ebert, "Spatial text visualization using automatic typographic maps," *IEEE Transactions on Visualization & Computer Graphics*, vol. 18, no. 12, pp. 2556–2564, 2012.

[5] A. Skupin, "The world of geography: Visualizing a knowledge domain with cartographic means," *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl 1, pp. 5274–5278, 2004.

[6] R. Brath and E. Banissi, "Evaluating lossiness and fidelity in information visualization," in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2015, pp. 93 970H–93 970H.

[7] M. Wattenberg and F. B. Viégas, "The word tree, an interactive visual concordance," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 14, no. 6, pp. 1221–1228, 2008.

[8] M. A. Hearst and D. Rosner, "Tag clouds: Data analysis tool or social signaller?" in *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008)*. IEEE, jan 2008, pp. 160–160. [Online]. Available: http://dl.acm.org/citation.cfm?id=1334515.1334989

[9] S. Bateman, C. Gutwin, and M. Nacenta, "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*. ACM, 2008, pp. 193–202.

[10] S. Lohmann, J. Ziegler, and L. Tetzlaff, "Comparison of tag cloud layouts: Task-related performance and visual exploration," in *Human-Computer Interaction–INTERACT 2009*. Springer, 2009, pp. 392–404.

[11] A. W. Rivadeneira, D. M. Gruen, M. J. Muller, and D. R. Millen, "Getting our head in the clouds: toward evaluation studies of tagclouds," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2007, pp. 995–998.

[12] A. Maps, "Typographic maps," http://www.axismaps.com/, 2015.

[13] M. Nacenta, U. Hinrichs, and S. Carpendale, "Fatfonts: combining the symbolic and visual aspects of numbers," in *Proceedings of the International Working Conference on Advanced Visual Interfaces*. ACM, 2012, pp. 407–414.

[14] E. Alexander and M. Gleicher, "Assessing topic representations for gist-forming," in *Proceedings of the International Working Conference on Advanced Visual Interfaces*. ACM, 2016, in press.

[15] B. Alper, H. Yang, E. Haber, and E. Kandogan, "Opinionblocks: Visualizing consumer reviews," in *Proc. of the IEEE Workshop on Interactive Visual Text Analytics for Decision Making*, 2011.

[16] C. Collins, F. B. Viégas, and M. Wattenberg, "Parallel tag clouds to explore and analyze faceted text corpora," in *Proc. of the IEEE Symp. on Visual Analytics Science and Technology (VAST)*, 2009.

[17] J. Haberman and D. Whitney, "Ensemble perception: Summarizing the scene and broadening the limits of visual processing," *From perception to consciousness: Searching with Anne Treisman*, pp. 339–349, 2012.

[18] J. Ross and D. C. Burr, "Vision senses number directly," *Journal of Vision*, vol. 10, no. 2, pp. 10–10, 2010.

[19] D. A. Szafir, S. Haroz, M. Gleicher, and S. Franconeri, "Four types of ensemble coding in data visualizations," *Journal of vision*, vol. 16, no. 5, pp. 11–11, 2016.

[20] M. Bostock, V. Ogievetsky, and J. Heer, "D^3 data-driven documents," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 17, no. 12, pp. 2301–2309, 2011.

[21] L. Ongaro, "jQCloud: jQuery plugin for drawing neat word clouds that actually look like clouds," https://github.com/lucaong/jQCloud, 2014.

[22] M. Davies, "Word frequency data from the corpus of contemporary american english (COCA)," http://www.wordfrequency.info, 2011.

[23] A. Kittur, E. Chi, and B. Suh, "Crowdsourcing user studies with mechanical turk," in *Proceedings of ACM CHI*. ACM, 2008, pp. 453–456.

[24] W. Mason and S. Suri, "Conducting behavioral research on amazons mechanical turk," *Behavior research methods*, pp. 1–23, 2011.

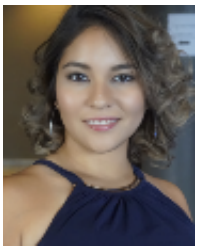
[25] L. R. Haber, R. N. Haber, and K. R. Furlin, "Word length and word shape as sources of information in reading," *Reading Research Quarterly*, pp. 165–189, 1983.

[26] M. Correll, E. Alexander, and M. Gleicher, "Quantity estimation in visualizations of tagged text," in *Proceedings of the 2013 ACM annual conference on Human Factors in Computing Systems*, ser. CHI '13. ACM, May 2013, pp. 2697–2706. [Online]. Available: http://graphics.cs.wisc.edu/Papers/2013/CAG13

[27] T. van der Geest and R. van Dongelen, "What is beautiful is useful-visual appeal and expected information quality," in *Professional Communication Conference, 2009. IPCC 2009. IEEE International*. IEEE, 2009, pp. 1–5.

Eric Alexander is an Assistant Professor of Computer Science at Carleton College in Northfield, Minnesota. He received his PhD in Computer Sciences at the University of Wisconsin-Madison in 2016. His work has primarily focused on the visual analysis of large collections of text, while his interests span information visualization, natural language processing, and the digital humanities.

Chih-Ching Chang is a PhD student supervised by Dr. Michael Gleicher at Department of Computer Sciences at University of Wisconsin-Madison. She received B.S. degree from Electrical Engineering and Computer Science department at National Chiao Tung University in 2014. Her research interest focuses on data visualization and human computer interaction.



Mariana Shimabukuro is a Masters of Science candidate from the Computer Science program supervised by Dr. Christopher Collins at University of Ontario Institute of Technology (UOIT). In 2013, during her undergraduate program in Brazil, she was granted a full scholarship from the Brazilian Government to do a 1 year exchange program at UOIT. Mariana obtained her Bachelors in Computer Science degree from Universidade Estadual Paulista (UNESP - Brazil) as top of her class in 2015. Her research interests fall into data visualization, HCI, recommendation systems, robotics and education.

Steven Franconeri Steve Franconeri is a Professor of Psychology at Northwestern, and Director of the Northwestern Cognitive Science Program. He studies visuospatial thinking and visual communication, across psychology, education, and information visualization.

Christopher Collins received the PhD degree from University of Toronto in 2010. He is currently the Canada Research Chair in Linguistic Information Visualization and Associate Professor at the University of Ontario Institute of Technology. His research focus combines information visualization and human-computer interaction with natural language processing. He is a member of the IEEE, a past member of the executive of the IEEE Visualization and Graphics Technical Committee and has served several roles on the IEEE VIS Conference Organizing Committee.

Michael Gleicher is a Professor in the Department of Computer Sciences at the University of Wisconsin, Madison. Prof. Gleicher is founder of the Department's Visual Computing Group. His research interests span the range of visual computing, including data visualization, image and video processing tools, virtual reality, and character animation techniques for films, games and robotics. Prior to joining the university, Prof. Gleicher was a researcher at The Autodesk Vision Technology Center and in Apple Computer's Advanced Technology Group. He earned his PhD in Computer Science from Carnegie Mellon University, and holds a B.S.E. in Electrical Engineering from Duke University. Prof. Gleicher is an ACM Distinguished Scientist.