Typeface size and weight and word location influence on relative size judgments in tag clouds

Article *in* Journal of Visual Languages & Computing · February 2018

DOI: 10.1016/j.jvlc.2017.11.009

Mirsad Hadzikadic
University of North Carolina at Charlotte

Mark Faust
University of North Carolina at Charlotte

Typeface size and weight and word location influence on relative size judgments in tag clouds

Khaldoon Dhou[a,*], Mirsad Hadzikadic[b], Mark Faust[c]

[a]*Department of Mathematics and Computer Science, University of Missouri – St. Louis*
[b]*Department of Software & Information Systems, University of North Carolina at Charlotte*
[c]*Department of Psychology, University of North Carolina at Charlotte*

Abstract

This paper focuses on viewers' perception of the relative size of words presented in tag clouds. Tag clouds are a type of visualization that displays the contents of a document as a cluster (cloud) of key words (tags) with frequency (importance) indicated by tag word features such as size or color, with variation of size within a tag cloud being the most common indicator of tag importance. Prior studies have shown that word size is the most influential factor of tag importance and tag memory. Systematic biases in relative size perception in tag clouds are therefore likely to have important implications for viewer understanding of tag cloud visualizations. Significant under- and over-perception of the relative size of tag words were observed, depending on the relative size ratio of the target tag words compared. The qualitative change in the direction of the estimation bias was predicted by a simple power-law model for size perception. This bias in relative size perception was modulated somewhat by a change to a bold typeface, but the typeface effect varied in a complex manner with the size and location of the tags. The results provide a first report of systematic biases in relative size judgment in tag clouds, suggest that, to a first approximation, simple power-law scaling models developed for simple displays containing 1-2 objects on a blank background, may be applicable to relative size judgments

*Corresponding author
Email addresses: dhouk@umsl.edu (Khaldoon Dhou), mirsad@uncc.edu (Mirsad Hadzikadic), mefaust@uncc.edu (Mark Faust)

in complex tag clouds. The results may provide useful design guidance for tag cloud designers.

Keywords: tag clouds, perception, psychophysics, size judgment

1. Introduction

A tag cloud is a visualization technique that is increasingly used to display the relative frequency of various keywords following a database or document search. It is a visualization of key search words (tags) in a cluster (cloud) where the importance of each keyword, as indexed by frequency in the search results, is indicated by key visual dimensions, e.g., typeface color, type, or size [1, 2]. Previous research indicates that size is the most prominent feature in a tag cloud [1, 2, 3, 4]. In a typical implementation of a tag cloud, typeface size communicates the frequency of words in a tag cloud and is a main variable controlled by tag cloud designers. However, the visualization literature indicates that size, as an indicator of quantitative information, has weaknesses due to the fact that words vary in shape in complex ways, and both typeface size and word length will affect the size of the imaginary text box used to present a word [1, 5]. This can cause some biases when viewers judge the relative size of words in a tag cloud as typeface size alone is typically used to index keyword frequency in tag cloud design. Moreover, the literature on viewers' perception of size indicates that size judgments of 2D regions (e.g., the rectangular text box region) themselves are systematically biased [6, 7].

Due to the complex mapping of keyword frequency onto typeface size [1, 3] typeface size in a tag cloud is an ambiguous indicator of keyword frequency. Relative typeface size, on the other hand, is monotonically related to relative frequency of keywords, and is therefore a perceptually salient visual feature available to the tag cloud viewer [5, 8]. Although relative typeface size is typically the design feature used to indicate the relative importance of keywords, a review of the existing peer-reviewed literature revealed only a single study investigating how users perceive the relative size of words in tag clouds [1]. We present more details of this study in the section covering related work, and note at this point that this study documented bias in relative size judgments in tag clouds, but did not control the length of tag words compared by viewers, nor did it focus on the systematic nature of judgment bias. The importance of relative size judgments to effective use of tag clouds, combined with the likelihood of biases in such judgments [1], suggests the importance of further study to document systematic biases in viewers of tag clouds. Because typeface size is often used in conjunction with other tag cloud design elements, it is important to study relative size judgment in the context of other common design elements [1, 3, 6, 9, 10, 11, 12].

The present study focuses on documenting systematic biases in relative size judgment in tag clouds while varying typeface weight and the location of the target tag word pair under comparison. We did so while maintaining a constant word length of target word pairs presented for comparison. We eliminated the effect of semantics by choosing non-sense and Latin words where we assume that the majority of participants are not familiar with Latin text. We also concentrated on the feasibility of applying a Steven's scaling approach, a general model for scaling perceptual biases [6], to the study of the perception of relative size of tags. Simple one-parameter psychophysical models of relative size judgment for 2 objects presented in isolation, based on a Steven's scaling framework, make specific testable predictions regarding the nature of systematic biases in relative magnitude judgment [7]. Demonstrating the applicability of this theoretical framework is an important first step towards developing a model that captures the size and scope of the biases in the perception of relative size in tag clouds.

2. Related Work

Tag clouds are popular navigational tools and their application has been investigated by researchers from different fields. One use of tag clouds is that they provide a summary of the relative importance based on frequency of appearance of keywords and users can use the tags to go to whatever block they want. This is observed on numerous websites, such as Flickr [13], which is a web application allowing people to host images and videos and Amazon [14], the well-known e-commerce company. Furthermore, tag clouds can be used to give users an impression of a content of a website or document. Tag clouds are even used in social websites to provide an impression of a person and his interests [15]. Tag clouds can also be used in interactive presentations. For instance, Hoeber and Liu [16] used tag clouds as one way to symbolize a search system. Their findings revealed the usefulness of interactive methods in the search method.

To study specific aspects of tag cloud usage, researchers have developed types of task to see how design elements affect different forms of tag cloud applications. For example, Rivadeneira et al. [17] proposed that tasks involving (a) impression formation of the topic present in the tag cloud, (b) allowing participant to browse and indicate what caught their interest, (c) searching for a keyword, and (d) tests of memory for keywords or topics, would be useful for study of the design elements on tag cloud use. They found that memory for specific keywords was affected by typeface size, with larger size leading to better memory. Similarly, Zhang et al. [4] explored the typeface size and location in Chinese tag clouds. They found that the recall of tags was influenced by their sizes and location. Likewise, Schrammel et al [18] examined the influence of size and layout. They had participants scan a tag cloud for 30 seconds, and found that participants were better able to recall words with a larger typeface size. Typeface size appears to have a more consistent influence on tag cloud memory, but aspects of location may be influential also. With regard to impression formation, Rivadeneira et al. [17] found that the ability to form an impression of the topic was influenced by typeface size and layout of the tag words. Felix et al. [1] used a couple of topic impression formation tasks. They found that a task where participants had a list of topic options to match against, the layout of the tag cloud affected task performance. They also found that an open-ended topic discovery task was only influenced by the size of the typeface of the tag words.

Search tasks are the most studied tag cloud task. Lohmann et al. [19] used three tasks: keyword search, search for the tag word with the largest typeface size, and search for tags related to a topic. Layout and typeface size influenced the results, but differentially for different search criteria. Keyword search was best for an alphabetical layout, size search was best for a layout where tags were spatially positioned by typeface size, and topic search was best for semantically clustered layout. Lohmann et al. [19] also tracked eye movements and found that eye movements were influenced by the layout of the tag cloud, but there was an overall bias for the upper left quadrant of a tag cloud to have the greatest amount of eye fixations. Halvey & Keane [2] found that country name search was affected by layout and typeface size, with larger size resulting in quicker search. Both layout and typeface font are important for tag cloud search tasks.

As discussed above, a recent study by Felix et al. [1] studied the influence of a range of tag cloud design elements (typeface size and intensity, layout of words in rows and columns versus a traditional tag cloud layout, use of geometric shading of varying sizes to indicate tag importance rather than typeface size) on tag cloud search and impression formation tasks. They found that different design elements had effects on the different tasks. The effect of layout and typeface size varied by task, but were influential when viewed across all of the tasks. Additionally, their study used a relative size ratio judgment task (Expt. 1) and found a statistically significant bias in ratio judgments. However, this study did not control the relative length of the smaller and larger target tag words under comparison and they were randomly selected from the tag cloud presented. This is important because the area of the text box for 2 different words of different sizes will change differentially as font size is changed. In the present study, we used matched pairs of test words that were of equal length when presented at the same font size. Felix et al. [1] also do not present any analyses looking at the systematic biases in the relative size judgments. This is the focus of the present study.

Tag clouds use a function of the frequency of keywords to determine how tags will be presented in the tag cloud. The typeface size is the primary visual characteristic in a tag cloud typically used to indicate the relative frequency of tags, however, tag cloud designers use several other types of presentation characteristics that could interact with the size to influence perceived relative size [1, 3]. For example, Bateman et al. [3] found that typeface size was the most effective indicator of tag importance, but typeface weight and color were also effective. Moreover, they found an interaction between the size of the word and the area of the tag text box. Dhou et al. [9] studied relative size judgments in tag clouds and found a systematic bias that systematically varied as a function of the typeface size of the larger comparison word (size of the smaller comparison word did not vary across task trials). One challenge to the study of perceptual biases in tag clouds is the wide variety of visual characteristics that are used to construct them, resulting in a large parameter space of potential study. Our approach to this problem is to employ a multi-factorial experimental design where multiple design characteristics are manipulated simultaneously to identify tag cloud design characteristics that individually, or in combination, impact relative tag size judgments.

To further pave the way to a better understanding of full scope of perceptual biases, we evaluate the feasibility of application of a Steven's scaling approach to developing psychophysical models to fit the relationship between actual and perceived relative tag size in future studies. Steven's scaling is a general approach to psychophysical scaling of perceptual magnitude that is well-studied and well-supported in the classic literature on human perception [6]. This framework models magnitude perception (e.g., size, loudness, brightness) as a power function of actual physical properties of a stimulus item. This approach has been used to study the effectiveness of graphical visualizations such as the systematic distortions in presentation size of geographic regions when projected onto a flat map [20], which necessarily increasingly distorts size for regions of increasing distance from the equator, and the perceived size of common graphical elements (e.g., bars in a bar chart, circles in a pie chart, [7]). Of particular interest to the current study, Spence [7] employed a single-parameter model of a relative size judgment task based on Steven's scaling assumptions. This simple single-parameter model predicts that relative size judgments are accurate when both objects being compared are nearly the same size, but are systematically increasingly biased as the true ratio of relative size increases. Figure 1 presents the qualitative predictions of this model, rescaled for direct ratio size judgments (e.g., target object is twice as large as the baseline object). As can be seen in Figure 1, the model predicts that when the power function exponent is less than one, relative size judgments are over-estimated for target objects that are smaller than the baseline object, and increasingly under-estimated as the target object becomes increasingly larger than the baseline object. Note that an exponent of one is simply the identity function for veridical relative size perception. Spence's [7] results indicated that perceived relative judgments of rectangular bars (e.g., included in a typical bar chart) could be fit with their model when the power function exponent was less than one (i.e., .76). Earlier studies of direct size estimation of rectangular regions indicate exponent ranges in the .75-.80 range [6, 20]. The current study will assume, based on prior studies of tag clouds [2, 3, 17], that, to a first approximation, the size of a word can be captured by considering the size of a rectangular region that captures the letters as presented.

As research indicates, prior work has looked at how design elements of tag clouds impact search, memory, and impression formation tasks, but given the common use of typeface size to indicate keyword importance, more work documenting the nature of relative size judgments in tag clouds is needed.

3. Hypotheses

In the present study, viewers judged the size of a target tag word that varied in typeface size from tag cloud to the next, in comparison to a baseline tag word that was held at a constant typeface size across presentations. We had the following predictions regarding biases in relative size judgments:

1. H1: We therefore predict that the perception of the relative size of tag words should be marked by both over-estimation of relative size when the target tag word is smaller than the baseline comparison tag, and an

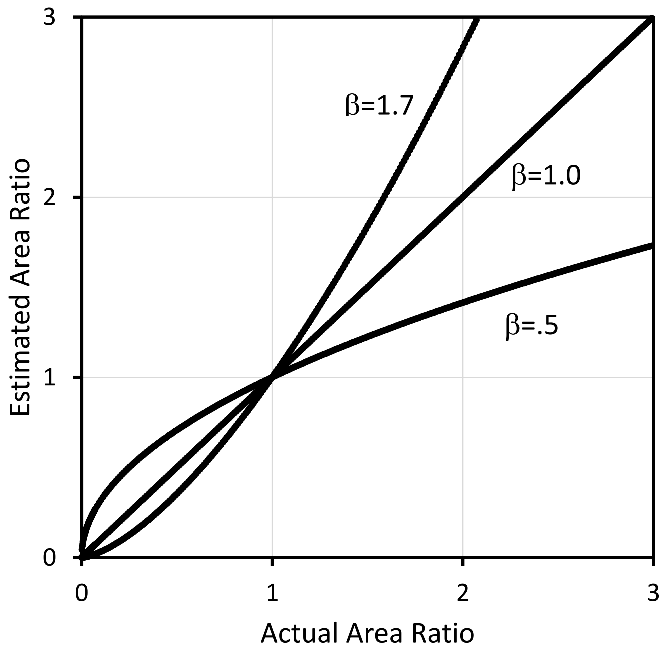

Figure 1: Predicted psychophysical models for a single-parameter ratio size judgment task, model based on Spence [7]. When the single parameter β is less than one, the model predicts ratio size judgements will be over-estimated when the test item is smaller than the reference item, and increasingly under-estimated when the test item is larger than the reference item. The model predicts an inflection perceptual bias when the test and reference items are about the same size. The model predicts the opposite pattern when beta is greater than one. With $\beta = 1$, the model predict accurate ratio size judgments across changes in relative size.

under-estimation at larger sizes, relative to the baseline tag. This is based directly on application of a single-parameter model of Steven's scaling and the prediction that word size judgments would yield results similar to previous studies of rectangular closed figures [6, 7, 20].

2. H2: As suggested by the results of Bateman et al. [3], who found that typeface weight (e.g., bold text) positively increased perceived tag importance, we predict that typeface weight will influence the perceptual bias of relative size judgments of tag words.
3. H3: Participants will overestimate the size of the words in the upper left

quadrant of the tag clouds. This is supported by the findings of Halvey and Keane [2] that information found in the upper left portion of the tag cloud or list is identified the fastest. Similarly, H3 is also supported by the study of Rivadeneira et al. [17] which found that the recall of the words in the upper left portion of the tag cloud was the highest. This seems logical given that in an occidental society people most often scan documents from left to right and from top to bottom [2]. The study of Zhang et al [4] also supports this hypothesis as they found that the recall of the tags in the upper left quadrant was significantly better than the recall of tags in the lower quadrants.

All of these hypotheses are evaluated in relation to our results in the general discussion section. The main goals of the study were to (a) map out for the first time the systematic biases in estimated size ratio in tag clouds as the size ratio of the tags being compared is varied, (b) test the feasibility of a simple single-parameter power law model (i.e., Steven's scaling model, [6, 7, 20] for predicting relative size estimation biases in complex tag cloud-style visualizations, and (c) evaluate the influence of other display dimensions (e.g., bold typeface, location in the tag cloud) on relative size estimation biases.

4. Experiment

This study aims to further our understanding of users' perception of estimated size ratio in tag clouds given the typeface weight (e.g. regular vs. bold) and the location of the words (e.g. location in a 4 quadrant grid).

4.1. Method

4.1.1. Participants

Data for this study was collected from a random sample ($N = 78$) of UNC Charlotte students. Because the study was conducted in a Western society, we assume participants were biased to read from left-to-right, and top-to-bottom. To further motivate performance, participants were informed that the most accurately answered questions would be awarded a $25 Starbucks gift card. Two of the participants entered the same answer for each question and thus, they were removed from the pool of participants. Two other participants reported that they did put random responses and thus, their responses were removed and not included in the analysis for the study ($N = 74$ for final sample analyzed). The remaining participants were 50 males and 24 females, age 18-50, $M = 22.4$ years. Participants reported having normal or corrected vision that allowed them to effectively read the text on the computer screen. The average time being spent on one screen for all participants was 6.8 seconds. The average time spent on the whole task was around 11 minutes per participant.

4.1.2. Materials

The experiment was conducted in a student computer with a proctor present. The lab was open in certain times and students could stop by and do the experiment. Each participant viewed 96 screens each of which had a question asking him to choose by how much one word (target word) is bigger than another word (anchor word). Target pairs of words in this experiment exist in a display of a tag cloud environment. The resolution of the screen was 800x600 pixels. These were standard flat screens in a certain size range 17-32 inches. The participants could spend as much time as they wanted on the task. The design involves a manipulation of 4 independent variables (IVs) and 2 control variables (CVs), as follows:

- Typeface weight: Half of the trials had words in a tag cloud in a regular typeface, while the rest of the trials had words in a bold typeface. All the words in the tag cloud have the same typeface weight manipulation as the target pair.

- Horizontal location: The location of the target pair varied between the left and right portions of the tag cloud.

- Vertical location: The location of the target pair varied between the upper and lower portions of the tag cloud.

- Typeface size: The baseline anchor word was always presented 18 pt typeface and the target word of the pair varied in size: 24, 30 and 36 pt.

In this experiment, two control variables (CVs) were used:

- Word pair: Two pairs of words were taken from a non-sense words website [21]. This was done to minimize bias from any semantic associations that meaningful words would have.

  First pair: dealizer (anchor word) vs. citounst (target word)

  Second pair: jarved (anchor word) vs. adring (target word)

  Figure 2 is an example of two screen shots where the pair of words is manipulated.

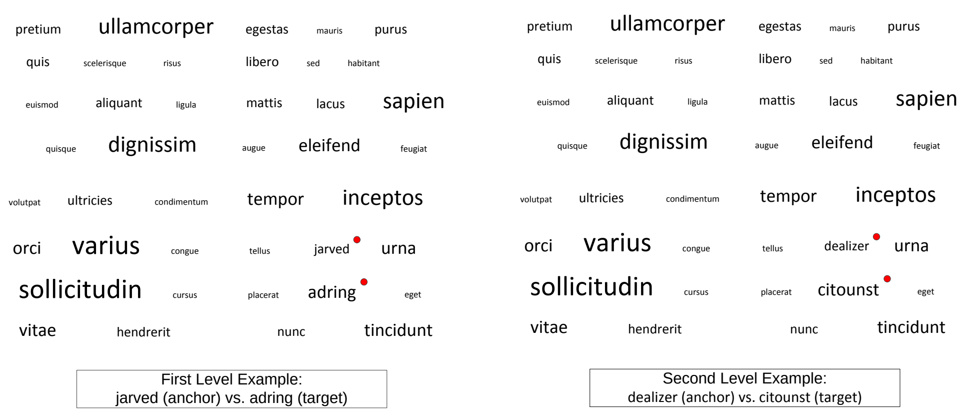

Figure 2: An example of manipulating pair of words CV used in the present experiment. In the left tag cloud the first pair of words (jarved vs. adring) was used while in the right tag cloud, the second pair of words was used (dealizer vs. citounst)

- Right/Left positioning: The right/left positioning of the target and anchor words of a tag word pair were swapped across trials.

The 4 independent variables and the 2 control variables were completely counterbalanced and added to the standard background tag cloud. This yielded distinct stimulus displays, or a 2 typeface weight (normal vs bold) x 2 horizontal location (right vs left) x 2 vertical location (top vs bottom) x 2 target typeface size (24, 30, 36) x 2 target location in pair (right vs left) x 2 target pairs factorial design with all manipulations within participants. The dependent variable was the estimated size ratio for the target pair of words in a tag cloud.

The two target pairs of words were selected according to the following criteria:

(a) In each pair, the two words must have the same total number of ascenders and descenders, which are the parts of the letter that extends above or below the level of the top or bottom of an x, respectively (i.e. b, p);

(b) In each pair, the two words must have roughly the same width when they are at the same size and when variable width typeface is used. For example, the two words in each target pair have almost the same width;

(c) In each pair, the two words must have the same number of letters. For example, the two words dealizer and citounst have 8 letters each

The rest of the words in the tag cloud are taken from *Lorem Ipsum*, that is a modified piece of Latin text used to fill layout designs. Latin text is assumed to be unfamiliar to most of participants, while still having the quality of forming pronounceable non-words in English. The reason to choose non-sense words and a Latin text is to make sure the effect of the semantics is eliminated because our interest is in how users judge the relative size of the words in tag a cloud without including any semantic effect. Words in the tag cloud (excluding the target pair) varied in their frequency. The frequency of the size of the words in the tag cloud is as in Table 1.

Table 1: The frequency of *Lorem Ipsum* words in the tag clouds used in this research

Typeface Size	**Frequency**
Typeface 12 pt.	17 times
Typeface 18 pt.	11 times
Typeface 24 pt.	6 times
Typeface 30 pt.	4 times
Typeface 36 pt.	2 times

Mirroring was performed in this experiment as a way to control because changing the words in the neighborhood of the target pair might add a bias to the judgment. The purpose of mirroring was to retain the words in the neighborhood of the target pair to avoid any bias in perception. Mirroring was performed as follows: First, dividing the tag cloud into 4 quadrants where there was a number of words in each quadrant and where the target and anchor words were located in one quadrant. The four quadrants were: Upper Left (UL), Upper Right (UR), Lower Left (LL) and Lower Right (LR). Second, varying the location of the word pairs by quadrant in the tag cloud. Figure 3 shows how the contents of the quadrants in a tag cloud change as the horizontal and vertical IVs are manipulated.

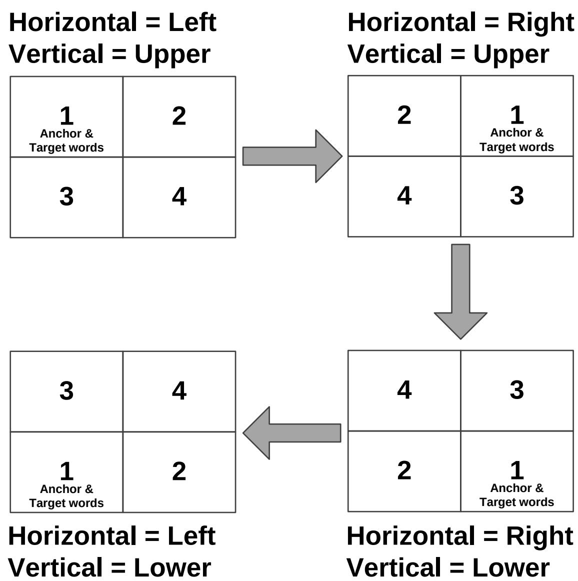

Figure 3: Changing vertical and horizontal locations of anchor and target words in a tag cloud. Every time the horizontal and vertical IV were manipulated, blocks change their location within a tag cloud.

Each relative typeface estimated size ratio at each particular relative size was compared with the typeface size ratio squared at that particular relative size. For example, the relative typeface estimated size ratio at relative size 18 vs. 24 was compared with the typeface size ratio squared at that relative size which is 1.778 ($(24/18)^2 = 1.778$). The reason this measurement was chosen is because when the typeface size gets doubled, both the height and width are doubled. This measurement is close to the area of the target word divided by the area of the anchor word, where the area is defined as the minimal box around the word [22].

4.1.3. Procedure

In the experiment, participants first completed an informed consent. Then, a participant was asked to review the experiment instructions where there was one example on how to perform a trial. The participants participated willingly. The experiment initiated with the display of 96 random displays, in which each of the displays was a trial. Each trial was different for each participant to lower the risk of participants copying from each other. For each display, the participants were asked to judge how big one word was in comparison with another. For each trial, the participant would choose an answer from a continuous scale that runs from 1 to 6 and hit an OK button to go to the next trial. Both the numeric and time responses (to the nearest tenth) were recorded. The screen remained visible until the participant recorded an answer and then a new display followed immediately. After the users finished the questionnaire, we asked them to provide a feedback about the experiment.

4.2. Results and Discussion

Estimated size ratio of words that had the same typeface weight, horizontal location, vertical location and relative size were averaged separately for each participant. The average estimated size ratios for each participant were submitted to a 2 typeface weight (regular vs. bold) by 2 horizontal location (left vs. right) by 2 vertical location (upper vs. lower) by 3 target typeface size (24, 30, 36) repeated measures ANOVA. The reason why we chose repeated measures design is because every participant participates in all the conditions used in the experiment. All the effects were reported as significant at $p < 0.05$. All mean values are reported in Table 2. The subsections below show the main effects that are significant and these are illustrated in Table 3.

Table 2: Mean values of size ratio judgment for the interaction effect typeface weight x horizontal x vertical x relative size in the experiment

Typeface Style	Horizontal Location	Vertical Location	Relative Size	Mean Estimated Size Ratio
Regular	Left	Upper	18 vs. 24	2.013
			18 vs. 30	2.803
			18 vs. 36	3.514
		Lower	18 vs. 24	2.013
			18 vs. 30	2.826
			18 vs. 36	3.575
	Right	Upper	18 vs. 24	1.979
			18 vs. 30	2.844
			18 vs. 36	3.427
		Lower	18 vs. 24	2.058
			18 vs. 30	2.798
			18 vs. 36	3.414
Bold	Left	Upper	18 vs. 24	1.988
			18 vs. 30	2.923
			18 vs. 36	3.521
		Lower	18 vs. 24	2.076
			18 vs. 30	2.872
			18 vs. 36	3.609
	Right	Upper	18 vs. 24	2.141
			18 vs. 30	2.823
			18 vs. 36	3.473
		Lower	18 vs. 24	2.088
			18 vs. 30	2.802
			18 vs. 36	3.621

4.2.1. The Effect of Typeface Weight

The main effect of typeface weight was significant, $F(1,73) = 9.294$, $p = 0.003$, $\eta^2 = 0.113$, indicating that the estimated size ratios of target pairs in bold typeface ($M = 2.828$) were significantly larger than the size ratio judgment of the target pairs in regular typeface ($M = 2.772$). This is consistent with the findings of Bateman et al. [3] that boldness has a strong visual effect in capturing the attention of the viewer.

4.2.2. The Effect of Actual Size Ratio

The main effect of actual size ratio was significant, $F(2,146) = 258.677$, $p < 0.001$, $\eta^2 = 0.780$. Figure 4 illustrates the participants' judgment of the

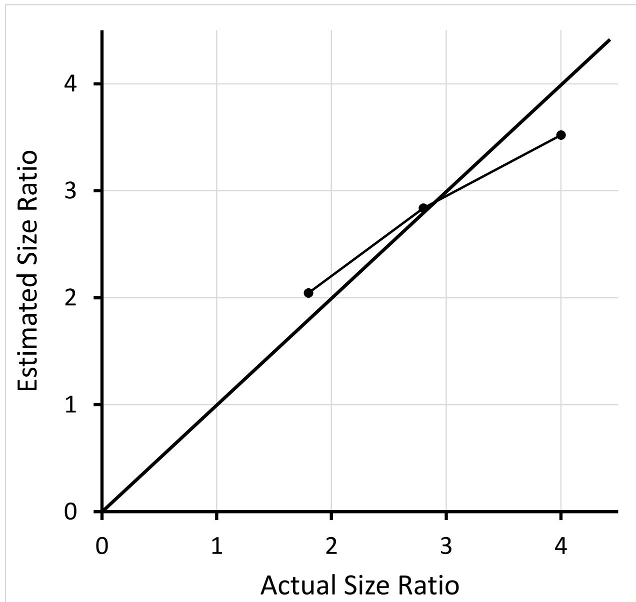

Figure 4: The main effect of actual size ratio plotted against the identity line indicating accurate size ratio estimation. Deviations from the identity line are interpreted in relation to the qualitative predictions of a single parameter Steven's scaling model with a search parameter less than one (see Figure 1). The horizontal axis is scaled by the ratio of typeface size squared (equivalent to the area ratio for a simple rectangular box model of word size): $(24/18)^2 = 1.778$ for relative size 18 vs. 24, $(30/18)^2 = 2.778$ for relative size 18 vs. 30 and $(36/18)^2 = 4$ for relative size 18 vs. 36.

three relative sizes used in the experiment. In addition, the predicted pattern of over-estimation of relative size for smaller size ratios, and under-estimation for larger size ratios was found, and the prediction that the crossover or inflection point would be when both target and anchor were about the same size (i.e., ratios of about 1) did not hold. Figure 4 also shows some non-linearity in that the under- and over-estimation effects are not the same size descriptively.

4.2.3. The Interaction Effect (Horizontal Location x Actual Size Ratio)

The interaction between horizontal location and actual size ratio was significant, $F(2, 146) = 4.783$, $p = 0.010$, $\eta^2 = 0.061$. Interpretation of this interaction

Table 3: Main effects and interactions that are significant in the experiment

Feature	**F value**	p
Weight	$F(1, 73) = 9.294$	$p = 0.003$
Size	$F(2, 146) = 258.677$	$p < 0.001$
Horizontal * Size	$F(2, 146) = 4.783$	$p = 0.010$
Weight * Horizontal * Size	$F(2, 146) = 4.818$	$p = 0.009$
Vertical * Size	$F(2, 146) = 4.755$	$p = 0.010$
Weight * Horizontal * Vertical * Size	$F(2, 146) = 3.389$	$p = 0.037$

is modified by the significant Typeface Weight x Horizontal Location x Actual Size Ratio interaction (see Table 3 and Figures 5 and 6).

Table 4: The interaction between horizontal and size

Horizontal	Relative Size	Mean	Typeface Ratio Squared
Left	18 vs. 24	2.022	1.778
	18 vs. 30	2.856	2.778
	18 vs. 36	3.555	4
Right	18 vs. 24	2.067	1.778
	18 vs. 30	2.817	2.778
	18 vs. 36	3.384	4

4.2.4. The Interaction Effect (Vertical x Size)

There was also a significant interaction between vertical and size, $F(2, 146) = 4.755$, $p = 0.010$, $\eta^2 = 0.061$. The pattern of cell means for this interaction are presented in Table 5. As can be seen, the significant interaction is driven by a greater under-estimation of the size ratio (at an actual size ratio of 4) for the test word pairs presented in the upper versus the lower half of the tag cloud. Comparison of mean estimated size ratios at the other actual size ratios (1.78 and 2.78, see Table 5) reveal exceedingly small differences between upper- and lower-presented test word pairs. These results run counter to assumption of a tendency for participants to have a top-to-bottom movement of an attentional window during viewing, which we predicted would lead to an opposite difference at the actual size ratio of 4.

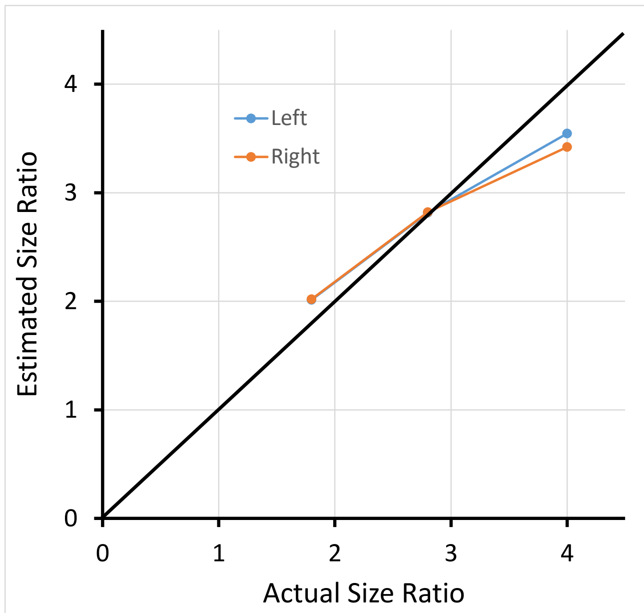

Figure 5: The interaction of horizontal location of the test word pair and the actual size ratio for regular typeface, plotted against the identity line indicating accurate size ratio estimation. Under-estimation bias is greater when the test word pair was presented in the right half of the tag cloud (orange line) compared to the left half of the tag cloud (blue line). The horizontal axis is scaled by the ratio of typeface size squared (equivalent to the area ratio for a simple rectangular box model of word size): $(24/18)^2 = 1.778$ for relative size 18 vs. 24, $(30/18)^2 = 2.778$ for relative size 18 vs. 30 and $(36/18)^2 = 4$ for relative size 18 vs. 36.

4.2.5. The Interaction Effect (Typeface Weight x Horizontal Location x Actual Size Ratio)

The interaction between typeface weight, horizontal and size was significant, $F(2, 146) = 4.818$, $p = 0.009$, $\eta^2 = 0.062$. This interaction was broken into two separate interactions at each level of typeface weight (regular vs. bold typeface). For each of these interactions, a 2x3 (Horizontal Location x Actual Size Ratio) repeated measures ANOVA was conducted at each level of the typeface weight. The interaction between horizontal and size was statistically significant for the regular typeface weight, $F(2, 146) = 4.2$, $p = 0.017$, $\eta^2 = 0.054$. Addi-

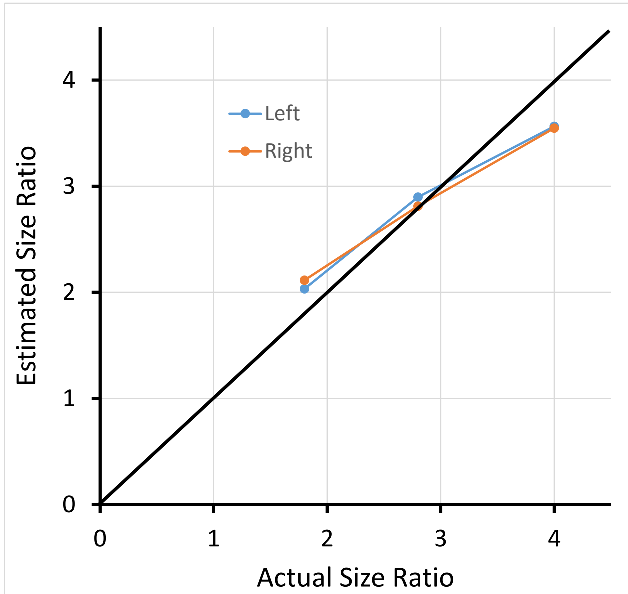

Figure 6: The interaction of horizontal location of the test word pair and the actual size ratio for bold typeface, plotted against the identity line indicating accurate size ratio estimation. For bold type face, estimation of size ratios is similar for test word pairs presented in either the right or left half of the tag cloud (see Figure 5 for comparison with regular typeface pairs). The horizontal axis is scaled by the ratio of typeface size squared (equivalent to the area ratio for a simple rectangular box model of word size): $(24/18)^2 = 1.778$ for relative size 18 vs. 24, $(30/18)^2 = 2.778$ for relative size 18 vs. 30 and $(36/18)^2 = 4$ for relative size 18 vs. 36.

tionally, for the bold typeface, the interaction between horizontal and size was statistically significant, $F(2, 146) = 5.411$, $p = 0.005$, $\eta^2 = 0.069$. These 2-way component interactions of horizontal location and actual size ratio are present in Figure 5 (regular typeface) and Figure 6 (bold typeface). As can be seen, the overall 3-way interaction (qualitative different patterns across Figures 5 and 6) is driven by a larger under-estimation of the size ratio for test word pairs presented in the right versus left half of the tag cloud for the regular typeface displays (Figure 5), but not for the bold typeface displays (Figure 6).

Table 5: The interaction between vertical location and actual size ratio

Vertical Location	Relative Typeface Size	Actual Size Ratio (Typeface Ratio Squared)	Mean Estimated Size Ratio
Upper	18 vs. 24	1.78	2.03
	18 vs. 30	2.78	2.82
	18 vs. 36	4.00	3.48
Lower	18 vs. 24	1.78	2.06
	18 vs. 30	2.78	2.82
	18 vs. 36	4.00	3.55

4.2.6. The Interaction Effect (Typeface Weight x Horizontal Location x Vertical Location x Actual Size Ratio)

The interaction between typeface weight, horizontal location, vertical location and actual size ratio was significant, $F(2, 146) = 3.38$, $p = 0.037$, $\eta^2 = 0.044$. However, examination of this complex 4-way interaction broken down in multiple ways does not reveal any need to present this interaction in detail as it does not affect interpretation of any of the lower order interactions that are the focus of our analyses. Moreover, the size of the component differences in mean size ratio estimates are small enough such that they most likely lack practical importance.

5. General Discussion

Our experiment quantitatively demonstrates systematic bias in the perception of relative size in tag clouds that ranged from an over-estimate of the larger tag by 14% for an actual size ratio of 1.78, to an under-estimate of 12% for an actual size ratio of 4.00. By contrast, moving from regular to bold typeface resulted in smaller changes in perceptual bias. Consideration of the interaction of typeface weight, horizontal location, and actual size ratio reveals that the use of bold typeface added 5% to the over-estimation bias for the larger tag at the 1.78 actual size ratio, and subtracted 3% from the under-estimation bias for the larger tag at the 4.00 actual size ratio. These maximal effects of bold (vs. regular) typeface were observed for test word pairs presented in the right half of the tag cloud. While not small, these moderate distortions of perceived relative size with actual size ratios in the range of 1.8-4.0 should assure tag cloud designers that serious large-scale distortions of perceived tag importance are unlikely for tag clouds that stay within this range of tag size ratios for the cloud as a whole. The use of bold text can modestly exacerbate the over-estimation of relative tag size, and by extension, relative tag importance, but this effect should be relegated to a relatively narrow range of relative sizes as the target tag approaches the comparator tag in size. On a more positive note, use of bold text for larger target tags, especially when presentation is in the right half of the tag cloud, may result in a modest reduction in under-estimation bias that appears at larger actual size ratios. Future studies will be needed to more fully map out the general shape of the psychophysical function for estimated relative size of tags for a greater range of typeface sizes. However, it should be recognized that power-law scaling (i.e., Steven's scaling, [6, 7, 20]) predicts that perception of size magnitude is generally increasingly under-estimated with increasing size of the object being viewed, and our results are suggestive of a similar increasingly severe under-estimation bias as tag size ratios climb over 4.

Our findings are consistent with those of Halvey and Keane [2] who emphasized the importance of typeface size in the design of tag clouds. Similarly, Bateman et al. [3] found a strong visual influence of the typeface size in tag clouds and determined that users could identify small variations in the typeface size. A major contribution of our study is that we provide the first evidence documenting systematic biases in the estimation of the relative size in tag clouds, thereby complementing previous work demonstrating the links between the size of tag words and perceived and remembered importance of tags [2, 3, 23, 24]. We also provide tentative evidence that power-law models of magnitude judgment (i.e., Steven's scaling, [6]), that have been effective in studies of the effectiveness of graphical visualizations [7, 20], can be extended to the perception of estimated size ratio in complex tag cloud visualizations. By assuming a simple rectangular text box model to index the size ratio of 2 words, we were able to extend the predictions of a simple one-parameter model of the estimation of the relative size of 2D rectangles [7], to successfully predict an inflection point in the direction of biased relative size estimation in tag clouds. That is, we successfully predicted that the estimation of the size ratio of 2 tags would be over-estimated at smaller size ratios, moving to under-estimation for larger size ratios. However, the simple-one parameter model (see Figure 1) predicts an inflection point for perceptual bias when the 2 areas under comparison are the same size. While we did not test tag size ratios in that range, we found an inflection point for the crossover of relative size estimation bias at size ratio of about 2.8. By employing a simple 2-parameter power-law scaling model used to predict size judgments of a single object [6, 20] we would be able to predict higher crossover points for under-to-overestimation of relative size, but at the loss of what seems to be a reasonable assumption that 2 tags of the same size would be correctly perceived to have the same size. One possibility is Steven's scaling, as it is based on simple power laws of the relationship between physical stimulus dimension magnitude and perceived magnitude on that dimension may be limited in its ability to fully capture the perception of the relative size of non-closed complex regions defined by text strings (i.e., words). However, our work suggests that a simple one-parameter power-law scaling model may capture biased estimation of relative size in tag clouds for the range of relative size above 1.7 or so. Further research should concentrate on examining the range of relative tag size ratios above 4 to provide a check of this model for the predicted large-scale distortions of estimated size ratio it predicts.

The fact that typeface weight had a significant effect on the bias of size ratio estimates is consistent with the work of Bateman et al. [3] who found that the typeface weight in a tag cloud had a strong visual effect. This confirms H2. Bold text was also a noticeable factor as some comments from participants about that were reported. Thirty nine of the users indicated that the typeface weight (i.e. bold vs. regular) influenced their perception of the relative size such as: "I based my decisions on the boldness of each of the typeface of each of the words and tried to visualize blowing them up from the small word to match the large", "The bold face letters led me to believe the word was bigger than it actually was" and "the boldness of the word creates the image of the word being bigger than it really is". Tag cloud designers should strongly consider the typeface weight to influence the attention of the viewers.

Based on the empirical results, it was also clear that for regular typeface, the left-right differences emerge at larger sizes for the target word, whereas, for bold typeface the left-right differences emerge for the medium and smaller target word sizes. These effects are small and may indicate different spatial attention biases for regular and bold typeface. Overall, bold and regular typefaces were not appreciably differentially biased to a degree that tag cloud designers should be concerned about this variable. Spatial attention may have small effects where relative word size ratio judgment is more distorted on the right than the left for both bold and regular typefaces, but in different size ranges. The lack of main effects for manipulation of horizontal or vertical location of the test word pair is generally inconsistent with the prediction (H3) of general left and top biases. However, the results do support a tendency for greater size ratio size estimates in the right side of tag clouds.

6. Conclusion

In this paper, we explored how the relative size of test word pairs of different typeface weights and locations in a tag cloud is perceived by viewers. We manipulated the typeface weight (regular vs. bold), the horizontal location (left vs. right), the vertical location (upper vs. lower) and the relative size of the 2 words. The main goals of the study were to (a) map out for the first time the systematic biases in estimated size ratio in tag clouds as the size ratio of the tags being compared is varied, (b) to test the feasibility of a simple single-parameter power law model (i.e., Steven's scaling model, [6, 7, 20]) for predicting relative size estimation biases in complex tag cloud-style visualizations, and (c) evaluating the influence of other display dimensions (e.g., bold typeface, location in the tag cloud) on relative size estimation biases. Our findings represent what we believe to be the first examination of biases in relative size estimation in tag clouds.

We report modest under- and over-estimates of tag size ratios in the 12-15% range for a range of tag size ratios in the 1.8-4.0 range, crossing over in direction near a tag size ratio of 2.8. The influence of bold, as opposed to regular, typeface was smaller and varied, being larger for tag word pairs presented in the right half of the tag cloud. These results provide important initial documentation of a sizable systematic bias in the judgment of the relative size of words in a tag cloud. The systematic nature of the bias presents an opportunity for future quantitative modeling and potential correction for this perceptual bias. Our results are consistent with the qualitative prediction of a simple single parameter power-law scaling model (extended from Spence [7]), which would predict our observation of regions of over-estimation and under-estimation of relative size. However, a second parameter will need to be added to capture a cross-over point that in not at unity ($\sim$2.8 in our results). Moreover, it seems reasonable to assume that the cross-over point for the size judgment bias (that marks the center of the low bias range) will vary with tag cloud design elements employed.

Future research should concentrate on mapping the bias across the tag cloud design space. Future work can also include eye-tracking studies that provide some visual patterns on the eye-fixation duration for tags with certain properties. In addition, research on estimated size ratios can be done for words in other languages such as Arabic, Hebrew and Chinese. This involves studying the properties of the letters in these languages and building the related mathematical models.

References

[1] C. Felix, S. Franconeri, E. Bertini, Taking word clouds apart: An empirical investigation of the design space for keyword summaries, IEEE Transactions on Visualization and Computer Graphics.

[2] M. J. Halvey, M. T. Keane, An assessment of tag presentation techniques,

in: Proceedings of the 16th international conference on World Wide Web, WWW '07, ACM, New York, NY, USA, 2007, pp. 1313–1314.

[3] S. Bateman, C. Gutwin, 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, HT '08, ACM, New York, NY, USA, 2008, pp. 193–202.

[4] Q. Zhang, W. Qu, L. Wang, How font size and tag location influence chinese perception of tag cloud?, in: Proceedings of the 9th international conference on Engineering psychology and cognitive ergonomics, EPCE'11, Springer-Verlag, Berlin, Heidelberg, 2011, pp. 273–282.
URL `http://dl.acm.org/citation.cfm?id=2029174.2029207`

[5] M. A. Hearst, Whats up with tag clouds?, Visual Business Intelligence Newsletter (2008) 1–5.

[6] R. Teghtsoonian, The standard model for perceived magnitude: A framework for (almost) everything known about it, The American journal of psychology 125 (2) (2012) 165–174.

[7] I. Spence, Visual psychophysics of simple graphical elements, Journal of Experimental Psychology: Human Performance and Perception 16 (1990) 683–692.

[8] J. Steele, N. Iliinsky, Designing Data Visualizations: Representing Informational Relationships, Oreilly and Associate Series, O'Reilly Media, 2011.
URL `http://books.google.com/books?id=Mp_R-vsOOEoC`

[9] K. Dhou, R. Kosara, M. Hadzikadic, M. Faust, Size judgment and comparison in tag clouds, poster presented at VisWeek 2013, October 13–18, Atlanta, GA (2013).

[10] K. Dhou, Toward a better understanding of viewers' perceptions of tag clouds: relative size judgment, Ph.D. thesis, University of North Carolina at Charlotte (2013).

[11] J. Emanuel, The millennials: Assessing the next generation of academic librarians, Ph.D. thesis, University of Missouri-Columbia (2012).

[12] K. Börner, Visualization framework & workflow design, available online from `http://ivmooc.cns.iu.edu` retrieved on August 2, 2013, lecture titled "Visualization Framework & Workflow Design" from Information Visualization MOOC (January 2013).

[13] Flikr, `http://www.flickr.com/photos/tags/` (retrieved Oct 2013).

[14] Amazon, `http://www.amazon.com/gp/tagging/cloud` (retrieved Oct 2013).

[15] S. Golder, B. A. Huberman, The structure of collaborative tagging systems, Journal of Information Science 32 (2005) 198–208.
URL `http://www.hpl.hp.com/research/idl/papers/tags/tags.pdf`

[16] O. Hoeber, H. Liu, Comparing tag clouds, term histograms, and term lists for enhancing personalized web search, in: Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on, Vol. 3, IEEE, 2010, pp. 309–313.

[17] A. W. Rivadeneira, D. M. Gruen, M. J. Muller, 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, CHI '07, ACM, New York, NY, USA, 2007, pp. 995–998.

[18] J. Schrammel, M. Leitner, M. Tscheligi, Semantically structured tag clouds: an empirical evaluation of clustered presentation approaches, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2009, pp. 2037–2040.

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

[20] S. E. Battersby, D. R. Montello, Area estimation of world regions and the projection of the global-scale cognitive map, Annals of the Association of American Geographers 99 (2) (2009) 273–291.

[21] Non-sense words, http://www.soybomb.com/tricks/words/ (May 2011).

[22] S. Bateman, personal correspondence by email (October 2013).

[23] I. Semetsky, Edusemiotics To Date, an Introduction of, Springer Singapore, Singapore, 2017, pp. 1–6. doi:10.1007/978-981-287-532-7_21-1. URL http://dx.doi.org/10.1007/978-981-287-532-7_21-1

[24] W. S. Cleveland, C. S. Harris, R. McGill, Judgments of circle sizes on statistical maps, Journal of the American Statistical Association 77 (379) (1982) 541–547.