EXPLORING THE EFFECT OF USER CHARACTERISTICS IN WORD CLOUD VISUALIZATIONS

ZEHRA SHIRIN

A THESIS SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF ARTS

GRADUATE PROGRAM IN INFORMATION SYSTEMS AND TECHNOLOGY
YORK UNIVERSITY
TORONTO, ONTARIO

MARCH 2022

Abstract

Word clouds are very popular for visually summarizing texts. While word clouds usually show the frequency of words using font size, recent studies have explored other possible design elements. However, there is still a gap in terms of understanding how individual differences among users may impact their performance. This thesis aims to bridge this gap by answering two key research questions: What user characteristics are impacted by different variations of word clouds, and how word clouds could be adapted to different user characteristics. To answer these questions, we ran a user study where participants performed *perceptual speed* and *verbal working memory* tests followed by 36 trials of the *magnitude judgement* task for word clouds. Results showed that user characteristics like *perceptual speed* can significantly impact the performance of users. These results can be useful in the future to provide personalized word clouds that are suitable for people with different user characteristics.

To Baba, Ammy,

Moiz,

& Muslim

Acknowledgements

I would like to praise and thank God, the almighty, who has granted countless blessings, knowledge, and the opportunity to complete this thesis.

Throughout the writing of this thesis I have received a great deal of support & assistance. I would like to thank my supervisor, Professor Enamul Hoque Prince, whose knowledge & expertise was invaluable in formulating the structure of this thesis. Your insightful feedback pushed me to sharpen my thinking and brought my work to a higher level. I would also like to thank Professor Manar Jammal for helping to supervise, and all other faculty & staff members of York University who helped me along the way.

I would like to thank my father, who I have always looked up to, and who inspired & guided me in the field of IT, among innumerable things. I would like to thank my mother for always pushing me to go further in my education, as well as in all aspects of life. These few sentences do not do justice towards a single thing you two have done for me. Countless trips to York University, endless hours of keeping me happy during my pregnancy, and watching my baby afterwards, are all just some of the things you have had to endure while I was working on this thesis. You have always been there for me.

I would like to thank my supportive husband Moiz, who never once doubted that I could pursue a master's degree or write a thesis, despite the challenges we faced as a newly-wed couple. I don't know where I would be if not for your encouragement & understanding.

I would like to thank my beautiful baby Muslim, who came into this world just as I was beginning to write this thesis and has perhaps been the most patient little man with his busy & overwhelmed mother. I hope this thesis inspires you to grow up with a passion for education & research.

I would also like to thank my siblings Hira, Azka and Ruhullah for putting up with me during some of my hardest times, my friend Alina for being with me through this journey and much more, as well as anyone else who helped in anyway during this time.

Thank you all, Zehra

Table of Contents

List of Tables

List of Figures

Introduction

1.1 Motivation

The advent of the World Wide Web and the subsequent rise of social media has led to rapidly growing text data. Given such exponential growth of text data, there is a need for retrieving information, summarizing the content, and browsing through large amounts of texts. The word cloud is an information visualization technique that focuses on showing the most frequently used words in a text document [43]. Typically, the word cloud uses different font sizes to indicate the popularity/frequency of different words, thus allowing the user to get a visual summary of the content.

Information visualization techniques use *graphical marks* such as bars, circles, and lines and *visual encoding channels* like length, area, and color of these marks [31] to represent data. A more non-typical mark, used specifically for textual information, is the text itself.

Figure 1.1: A word cloud made on the D3 word cloud generator by Jason Davies [12][48], using text from a CNN news report on the COVID-19 pandemic [18]

Text visualization is an active research area where many methods have been proposed to help analyze textual data [2] such as word clouds [43], visualizing document concordances and word frequencies [46], and visualizing literature and citation relationships [19]. Among them, word clouds have been popular, having appeared on various news, social media, and blogging sites. Word cloud visualizations usually show the frequency of words using *font size*. However, recent studies have explored other possible visual encoding techniques such as using colors or bars to encode word frequency as well as layouts such as row-wise and column-wise organizations and semantic groupings [14, 20]. As can be seen in Figure 1.1, there is a rich design space for word clouds that can use hues, font faces, font sizes, angles, spatial layout, etc., for encoding information or for styling purposes.

Even though the studies on word cloud designs have evaluated various alternative visual encoding methods and layout types [14, 20], they generally do not consider individual differences among users or explore how to support users with a wide variety of backgrounds and abilities. However, recent research in visualization suggests that individual user needs and abilities can have a significant impact on their performance when using visualizations. Previous work has been done on individual differences and their impact on task performance for commonly used charts such as bar charts [8, 11, 10, 25]. However, there exists a gap in such research when it comes to text visualizations, specifically word clouds. For example, existing works experimented with the design space of word clouds [14, 20]; however, they do not test or make any conclusions based on the individual differences of the users completing the word cloud-based tasks. This thesis aims to bridge this gap by examining what user characteristics are impacted by different variations of word clouds, and how word clouds could be adapted for different user characteristics.

1.2 Approach

To bridge the gap discussed above, this thesis aims to answer the following research questions:

1. What to adapt: What user characteristics are impacted by different design variations of word clouds? This thesis aims to answer this question by evaluating multiple measures of task performance (such as speed and accuracy) of users by testing them (using word clouds) against two different user characteristics, *perceptual speed* and *verbal working memory*.

Perceptual speed measures how fast someone can compare differences or similarities in a set of things, and *verbal working memory* refers to one's memory span for temporarily remembering verbal information. By answering this question we will be able to learn which specific user characteristics are impacted by specific adaptations to word cloud designs.

2. How to adapt: How can word clouds be improved to support users with specific characteristics? This thesis aims to answer this question by analyzing how users with specific user characteristics (e.g. low perceptual speed) may find certain designs of word cloud to be more or less effective than others. By answering this question we hope to contribute to future design interventions and support by providing design guidelines concerning different user characteristics.

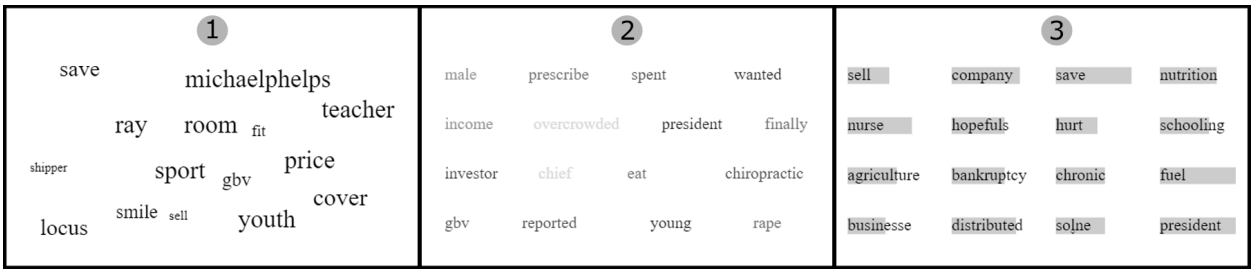


Figure 1.2: Examples of word clouds using different layout and visual encoding methods: (1) spatial with font size, (2) row with color intensity, (3) column with bar length

To answer the above research questions we design a user study pertaining to a specific task type. The study involves word clouds made from a combination of different types of *value encoding* methods and *layout* types. Value encoding refers to encoding data into visualizations by using a combination of graphical marks (text and bars) and their visual channels such

as color intensity, font size, angles, position, area, length, etc., while layout refers to the arrangement of the words in the word cloud such as row, column, or spatial (with no specific pattern). Figure 1.2 shows some examples of layouts and visual encoding methods.

Lastly, to bring our study together, the task we focus on is *magnitude judgment*, where the participants perceive/judge the differences between the frequencies of two words based on the visual encoding (e.g. font size, color intensity). This task was elected from a shortlist of other tasks, which will be discussed in Chapter 2.

1.3 Research Contributions

Previously word clouds have been tested for their design space and performance based on different tasks. However, to the best of our knowledge, none of these works focused on understanding the influence of individual differences over different layouts and visual encoding techniques for word clouds. This thesis analyzes word clouds from a new angle and bridges the gap, between research on *the word cloud design space* and work on *user adaptive information visualizations based on user characteristics*, by studying word cloud design in relation to individual differences and proposing design interventions specific to user characteristics.

Our research contributions relate to the answer of our two research questions, *what to adapt* and *how to adapt*. By analyzing the results of our study we found that there was a strong relation between perceptual speed and the performance of users in the magnitude judgment

task. We observed that perceptual speed significantly impacted the user performance across three different measures (based on errors made by users and task completion time). We also found that those with *low perceptual speed* performed just as well as those with *average perceptual speed* when using the *row* layout or the *color intensity* visual encoding method. These results can be used to provide personalized word clouds for those with lower measures of user characteristics.

1.4 Thesis Outline

The rest of the thesis is organized as follows. In Chapter 2, we present a literature review of work related to the word cloud design space and the impact of user characteristics on perceiving visualizations. Next, in Chapter 3, we describe the details of the user study including the word cloud designs that we experiment with, the user characteristics that we test as well as the tasks and procedure of the study. Then, Chapter 4 presents the results of the user study in details that help us answer our key research questions. Finally, Chapter 5 concludes the thesis by summarizing the key findings of this thesis as well as several directions for future work.

Literature Review

This literature review consists of three main sections. Section 2.1 discusses prior work on text visualizations with a focus on word clouds. Next, Section 2.2 gives an overview of user characteristics that may affect the way people understand visualizations. Finally, Section 2.3 describes the types of tasks that can be used to study the impact of user characteristics on the interpretation of word clouds.

2.1 Word Clouds

In this section, we will provide an overview of word clouds, the design space of word clouds, and an overview of word cloud algorithms. Interested readers are also directed to the work of Torres Paerjo et al. [43] for a detailed survey on word clouds.

2.1.1 Word Clouds

Word clouds are a text visualization technique that show the most frequent words in a text. Often these visualizations appear with the shape of a cloud with different font sizes indicating the frequency of the word. People may also refer to them by the term 'tag cloud', however in a tag could, words are assigned by users instead of extracting from text [43].

Word clouds can be used for a variety of text analysis tasks and have been described as ranging between being social (emotional) and being analytical [45]. They are considered for social tasks in the sense that they are used for blogging and news websites as a means to attract readers and give them an idea of what the piece of writing involves, while they are considered for analytical tasks in the sense that they are used for more technical tasks such as summarizing the frequencies of keywords in a document.

When word clouds are compared to other forms of data visualizations (such as bar charts) they have been found to be less efficient in tasks that involve judging the frequency of the words, while being more efficient in tasks related to extracting main concepts (regardless of frequency) and summarizing topics [20, 14]. However, data visualization methods such as bar charts also have their limitations when compared to word clouds. For example, bar charts can be less space-efficient compared to word clouds as they use bars to represent word frequency while in a typical word cloud, the word itself encodes the frequency/popularity using font size.

2.1.2 Design Space of Word Clouds

Word clouds offer numerous ways of styling to help the reader better understand the topics being summarized in the word clouds. Word clouds can be manipulated using font types, font sizes, angle of the words displayed, spacing between the words, groupings of related words, color, etc. Studies have concluded that organizing words into groups or "zones" by their meaning, and displaying the zones in visually distinct groups, via spatial or color grouping, may not be aesthetically pleasing but is more efficient for analytical tasks [20]. Seminal studies like the one by Rivadeneira et al. [35], have researched specific tasks such as impression formation, and have concluded guidelines to constructing effective word clouds by using specific font features (font weight, font size, font color) as well as word placement (sorting, clustering, spatial layout). Overall, Halvey et al. [17] and Rivadeneira et al. [35] found that styling word clouds can greatly impact their effectiveness. Other than font size and positioning of words, which are a very basic properties of a word cloud, properties such as alphabetization have also been concluded in helping readers find specific information within word clouds [17].

A more recent study from Felix et al. [14] explored the design space of word clouds based on four different visual encoding methods (e.g. font size, color intensity, bar length, circle area) and three different layouts (row, column, and spatial). They found that the effectiveness of a word cloud design may depend on the task the user is performing. We provide more details on various possible user tasks with the word cloud in Section 2.3.

Beyond word clouds, researchers have also examined the usefulness of styling other forms of visualizations such as bar charts. For example, Carenini et al. found in a study [8] that bar graphs can be more effective when highlighting interventions are provided by bolding the outlines of bars or de-emphasizing irrelevant bars when the user is performing a task.

2.1.3 Keywords vs. Keyphrases

Word clouds are commonly criticized for not being effective with analytical tasks such as summarizing topics and extracting themes. Therefore, some researchers explored a variation of word cloud that shows keyphrases instead of individual words. The idea is that rather than having vague and separated words such as "sports" or "cars", a keyphrase such as "sports cars" may better describe the text in the visualization. This is because putting words into phrases allows the reader to understand the context of those words based on how they were used in their text of origin. Hearst and Rosner [21] imply that using keyphrases rather than keywords may also help word clouds to become more known for analytical tasks rather than a form of social visualization (a social marker; perceived as being fun, popular or hip).

Figure 2.1: Phrase-based word cloud made by Neil Turner [44]

Figure 2.1 shows a phrase-based word cloud made by Neil Turner, a UK-based product designer and researcher, who explains that *Wordle* word cloud generator was used with advanced features along with tilde symbols (‘∼’) to preserve phrases.

In the area of natural language processing various phrase extraction methods have been proposed that aim to find the important phrases based on a reference corpus, external taxonomies, or genre-specific document structures [26, 32, 9]. Chuang et al. develop a method for grouping similar words and varying the specificity of displayed phrases so that the word cloud can show selected keyphrases dynamically based on the available screen space and current context of interaction [9]. They conclude that subjects primarily choose multi-word phrases, preferred terms with medium commonness, and largely used phrases already present in a document. Phrase Clouds have great research potential, and their effectiveness, specifically on users with varying cognitive abilities, can be researched comparatively with basic word clouds.

2.1.4 Word Cloud Algorithms

Research has been done to make or improve word cloud algorithms such as the work by Johann Schrammel [36], or the word cloud class for the D3 JavaScript library written by Jason Davies [12]. Wordle is known for making graphic statements and some work, like the one by Viegas et al. [45], has been done on the design and usage of Wordle and how it stands out from other online tools such as TagCrowd [40]. Other online tools include MonkeyLearn's WordCloud Generator (which uses a stemming process to automatically detect collocations) [50], WordClouds.com [51], WordItOut (known for supporting emojis and other symbols) [52], WordArt (artistic tool that allows users to apply shapes to the word clouds) [49], and Abcya (an educational tool for children) [1].

2.2 Impact of User Characteristics in Interpreting Visualizations

Individual differences, such as cognitive abilities, personality traits, and other skills, have been proven to affect the way a person perceives and understands a visualization [8, 11, 10]. Cognitive abilities include need for cognition, perceptual speed, verbal working memory, and visual working memory. Personality traits include internal or external locus of control, as well as introvert vs. extrovert, while literacy and skills include visual literacy, expertise,

frequency, and reading proficiency. Table 2.1 lists different user characteristics along with their definitions. We discuss these characteristics below.

User Characteristic	**Definition**
Visual Literacy	The ability to interpret visualized information.
Reading Proficiency	The vocabulary size and ability to comprehend text read in English.
Verbal IQ	The measure of an individual's overall intellectual abilities.
Need for Cognition	The tendency or motivation to engage in cognitive activity, where persons with a high need for cognition act after careful assessment of information [3, 7, 25].
Perceptual Speed	The ability to compare or recognize symbols or figures and other simple tasks.
Verbal Working Memory	The amount of verbal information that can be maintained temporarily while activities or manipulations are performed on this memory [5].
Visual Working Memory	The amount of visual information that can be maintained temporarily while activities or manipulations are performed on this memory [28].
Spatial Visualization	The ability to mentally manipulate (rotate, cut, fold) figures.
Spatial Memory	The ability to remember configurations, locations, and orientations of objects.
Disembodiment	The ability to hold perceptual material in mind while disembodying it from other material.

Table 2.1: User Characteristics and their Definitions

2.2.1 Cognitive Abilities

Previous research found that various cognitive abilities impact the performance and subjective impressions of users when they interact with information visualizations. When working with bar graphs, Carenini et al. [8] and Conati et al. [11] tested perceptual speed, verbal working memory and visual working memory. Carenini et al. [8], in line with previous findings, found a relation between the bolding intervention (highlighting bars in a bar chart related to the given task) and the performance of users that have an average visual working memory. They also found that participants with higher cognitive abilities performed significantly better with tasks that had greater complexity. In another work, Conati et al. [11] found that users with lower measures of perceptual speed performed worse than others with higher measures in almost all tasks, while those with lower measures of verbal working memory performed worse in sorting-related tasks. With these findings, they concluded that while this signals a need for personalized help, there is still room to discover which specific features of a visualization affect performance if a user has a lower measure of a specific user characteristic.

Conati et al. [10] chose to delve deeper and include Spatial Visualization, Disembodiment, Need for Cognition, as well as different learning styles (such as Active vs. Reflective, Sensing vs. Intuitive, Visual vs. Verbal, and Sequential vs. Global), along with perceptual speed and visual working memory, when researching Georgia Basin Quest (GB-Quest) and QuestVis for interpreting geographical data. They found that in one of the task types (which involved comparing how the values of two sets of variables change as a whole), users with high

perceptual speed performed the task better if they could see data via colored boxes, while users with low perceptual speed performed better with a radar graph. They concluded that (if the task involves comparing overall changes in variables of interest) the system should be adaptive in selecting and recommending the visualization type based on the users' perceptual speed. (Perceptual speed would be tested for and fed to the system before the user uses the system.)

2.2.2 Personality Traits

A common personality trait that was studied is the locus of control. Locus of control is described as how a person perceives the amount of control they have over occurrences in their lives and can be characterized as either internal or external locus of control, implying that the person's behavior is in response to internal or external circumstances, and life outcomes are a result of either the person's own actions and abilities or from external factors out of their control accordingly [4]. One example of research involving locus of control and visualizations is the one by Ziemkiewicz et al. [53], which uses four different visualizations to study how locus of control influences an individual's use of a complex visualization system and finds that overall participants with an external locus of control were more accurate when answering questions as compared to those with an average or internal locus of control.

2.2.3 Literacy and Skills

While testing users with expertise in bar charts, Conati et al. [11] and Carenini et al. [8] classified expertise as basic or complex and found that expertise affects the performance of the user in low-level tasks. Conati et al. [11] concluded that this signals there should be personalized support available to non-experts when performing complex tasks.

When working with Magazine-Style Narrative Visualizations (MSNV) Lallé et al. [25] chose to use visual literacy, reading proficiency, and verbal IQ, along with the cognitive abilities of verbal working memory and need for cognition. The results of this study showed that users with low levels of visual literacy benefited from interventions as they have a lower ability in understanding data visualizations.

2.3 User Tasks for Word Clouds

For measuring the performance of users with visualizations various low-level tasks such as retrieve value, find extremum, sort, compute derived value, filter, characterize distribution, etc. were used by Schrammel et al. [37], Conati et al. [10], Halvey et al. [17], and Rivadeneira et al. [35].

A study that studied word clouds and their design space more specifically is the one by Felix et al. [14]. It focuses on four main tasks: Magnitude Judgement (extracting quantitative

information), Keyword Search, Topic Matching (matching a word cloud to predefined topics), and Topic Discovery (extracting a topic from the word cloud). In the magnitude judgment task, this study concluded improved accuracy when the word cloud had additional marks (bars and circles with size based on frequency) and lower performance linked to *spatial layout* (in contrast to *row* and *column layout*). In the keyword search task it found improved performance time based on font properties, *spatial* and *column layout*, and quadrant while having no major findings in the *topic matching* and *topic discovery* tasks.

In comparison to analytical tasks, when it comes to conducting search tasks, word clouds have been criticized, and at times proven to be a weak form of text visualization. Sinclair and Cardew-Hall [38] concluded that general information-seeking tasks were more compatible with word clouds, as compared to specific information-seeking tasks where traditional search interfaces were more preferred. Kuo et al. [24] found word clouds to be less effective in identifying relational concepts and concluded that they tend to slow users down when trying to retrieve specific items.

User Study

The goal of the user study is to find the answer to the two research questions presented: *What user characteristics are impacted by different design variations of word clouds?* and *How can word clouds be improved to support users with specific characteristics?*

To answer the above questions we designed a user study where participants were required to go through a series of tests to measure their user characteristics and then were asked to complete a task using the different word clouds designs provided. We then analyzed the results and performed statistical tests to understand the potential impact of user characteristics on user performance and to explore the adaptability of the word clouds based on specific user characteristics. The following sections of this chapter introduce the user characteristics that we included in the study, the word cloud designs that we evaluated, and the detailed study procedure and task setup.

3.1 User Characteristics

In the user study, participants went through two standard tests of psychology for user characteristics namely *perceptual speed* and *verbal working memory*. *Perceptual speed* is described as "the ability to compare or recognize symbols or figures and other simple tasks" [47], while *verbal working memory* refers to the amount of verbal information that can be maintained temporarily while activities or manipulations are performed on this memory [28, 5].

The above-described user characteristics have been shortlisted from a long list (as can be seen in Table 2.1). These characteristics relate directly to the ability to read visualizations as they have been previously tested by researchers with regards to other similar visualization-based studies and have been found to have an impact on performing tasks [14, 41]. Although many other characteristics could be tested, a small number has been selected so that the participants do not take too much time while answering these tests before the study. This way, we keep the total duration of each study session to be reasonable.

We did not perform any standard English proficiency or reading comprehension tests. Instead, since we ran the study using Amazon Mechanical Turk (MTurk), we set up qualification criteria by choosing participants who consistently demonstrated a high degree of success in performing a wide range of HITs (Human Intelligence Tasks) [30]. This MTurk qualification met the need for the participants of the study to have basic English proficiency

and reading comprehension.

3.1.1 Perceptual Speed Test

Perceptual speed was calculated using an Identical Figures Test (P-3) [13], which was originally designed to be used as a paper-based test, and has been used so in many prior studies. For our study, we opted to digitize the test to make the study process seamless. As the paper-based test was 96 questions divided into 4 pages, the digital version was also made with 4 HTML pages (along with another instruction page) that appeared to the participants during the study.

Figure 3.1: An example of the questions in the Identical Figures Test (P-3). The participant must select one of the five figures on the right to match the left-most figure.

The questions were from scanned copies of the original graphics from the paper-based test, formatted along with radio buttons to make selection as easy as possible. See Figure 3.1. Participants were given 25 seconds to complete each page (they would automatically be taken to the next page once the time for that page was up). As the time taken on paper differs slightly from using a mouse to select the answers on the screen, the time limit given to the participants was based upon a pilot study. The score was the number of correctly answered questions within the time limit (with a range of 0-96).

3.1.2 Verbal Working Memory Test

Verbal working memory was tested using the Operation-Word Span Test (OSPAN). This test is a Java program authored by Taiyu Lin at ALTRC, Massey University, New Zealand [27]. The test shows a series of equations that the participants must classify as true or false, and shows the participants some words between each of the equations. The participants are asked to recall the words that they were shown in-between each equation. Figure 3.2 shows the general flow of the test.

The test works by making the participants recall information that was unattended for a small amount of time (while they were solving math equations)[23], and thus gives them a score ranging from 0 to 6 based on their average word span in working memory [41].

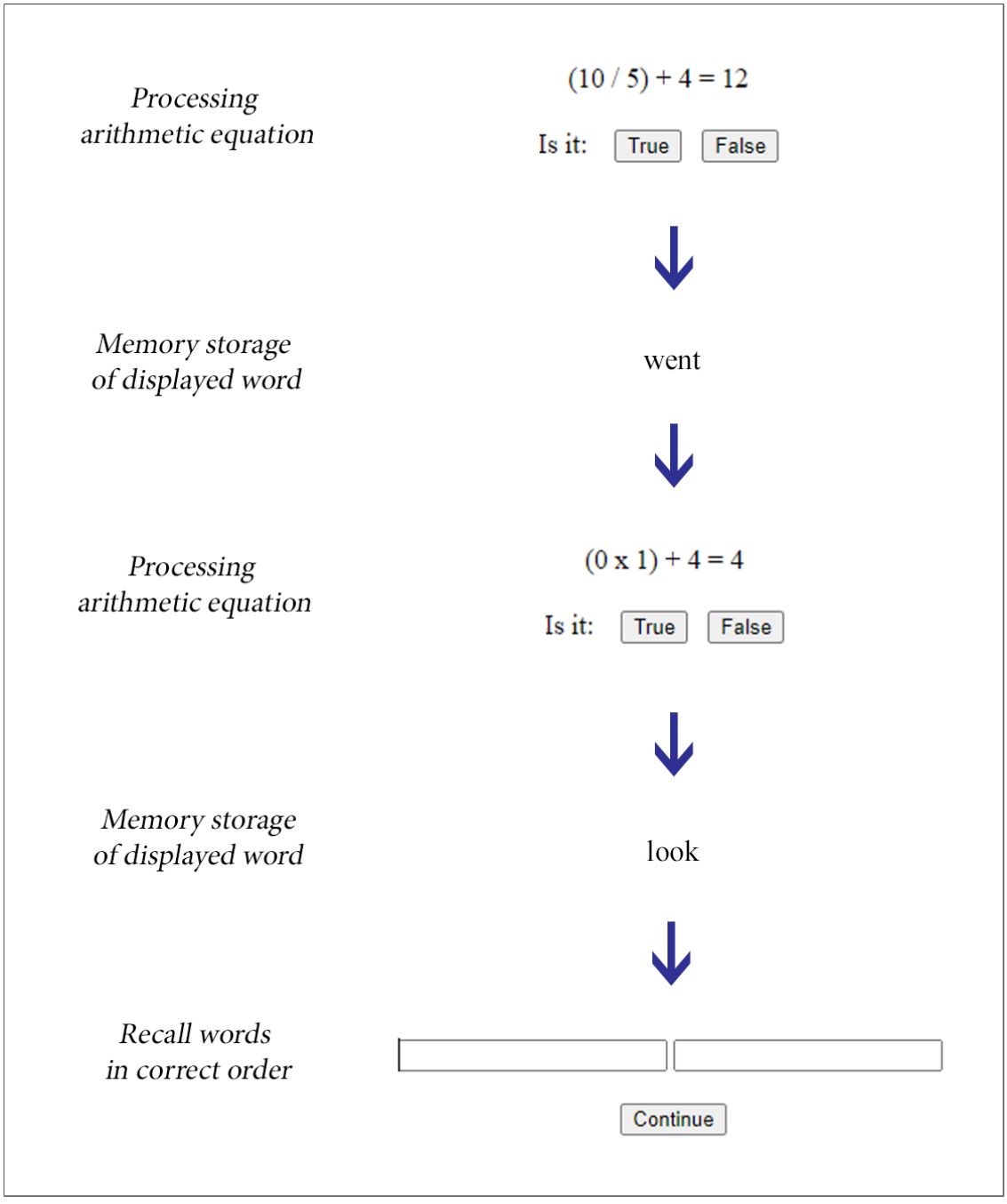


Figure 3.2: An example of the steps in the verbal working memory test (During the study each step is displayed on a separate page for the participant)

3.2 Task: Magnitude Judgement

In this task, participants were asked to judge the frequency of a given word based on its surrounding words. Users were shown word clouds containing 24-25 words, where two of the words in the word cloud had a red arrow pointing to them. The users were then asked to compare the words with each other and guess how much more frequent one word was from the other based on the visual encoding method used in the word cloud. Figure 3.3 shows an example of the task including the question, the word cloud, and the timer. The multiple-choice options listed four percentages that were from a list of 10 possible proportions of frequency. Section 3.4 further discusses the origin of the data and how it was used.

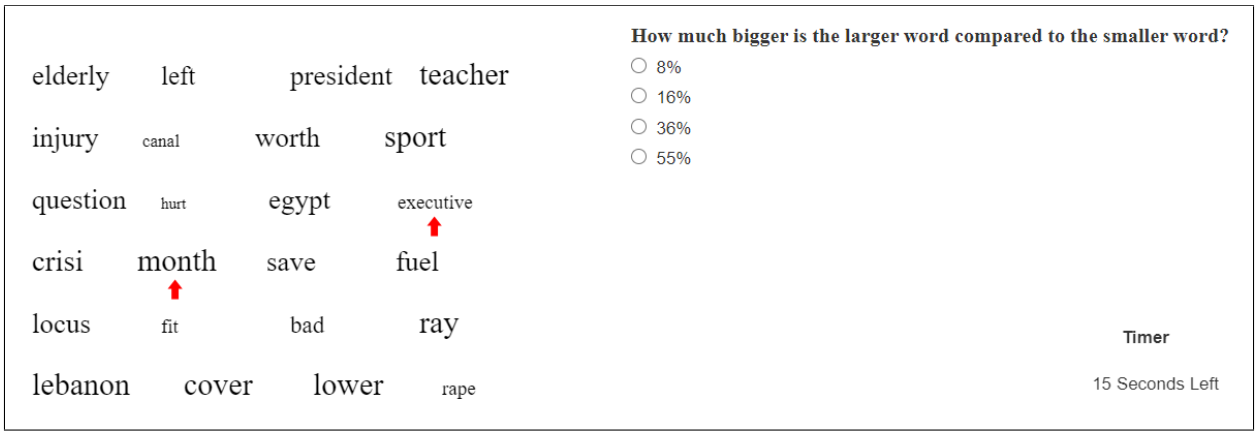


Figure 3.3: An example of a trial for the task of Magnitude Judgement which includes the word cloud (using the *row* layout and *font size* visual encoding method), the multiple-choice question, and a timer. The red arrows in the word cloud point to the two words that the question refers to.

This task was chosen from a list of tasks [14] that also included *Topic Discovery* (finding

the topic of the word cloud based on the words in it), *Topic Matching* (matching a topic to the word cloud from a given list) and *Keyword Search* (finding specific words in the word cloud and identifying them by selecting which quadrant they are in). The tasks of *Topic Discovery* and *Topic Matching* highly resemble each other. The task of *magnitude judgment* was chosen over these two tasks as *magnitude judgment* directly involves judging the frequency of words, something word clouds are most commonly used for.

The task of *Keyword Search* was also not selected for this study as research shows that word clouds are not as effective as other forms of text visualizations when it comes to searching words. As discussed in the literature review it has been previously concluded by researchers that general information-seeking tasks were more compatible with word clouds, as compared to specific information-seeking tasks [38] and word clouds are found to have slowed users down when in trying to retrieve specific items [24].

3.3 Design Space of Word Clouds

The study involves a combination of *layout* types and *visual encoding* methods. Figure 3.4 shows examples of the word clouds created from the possible combinations. The number of layouts and types of value encoding have been limited to allow for the completion of the study in a timely manner.

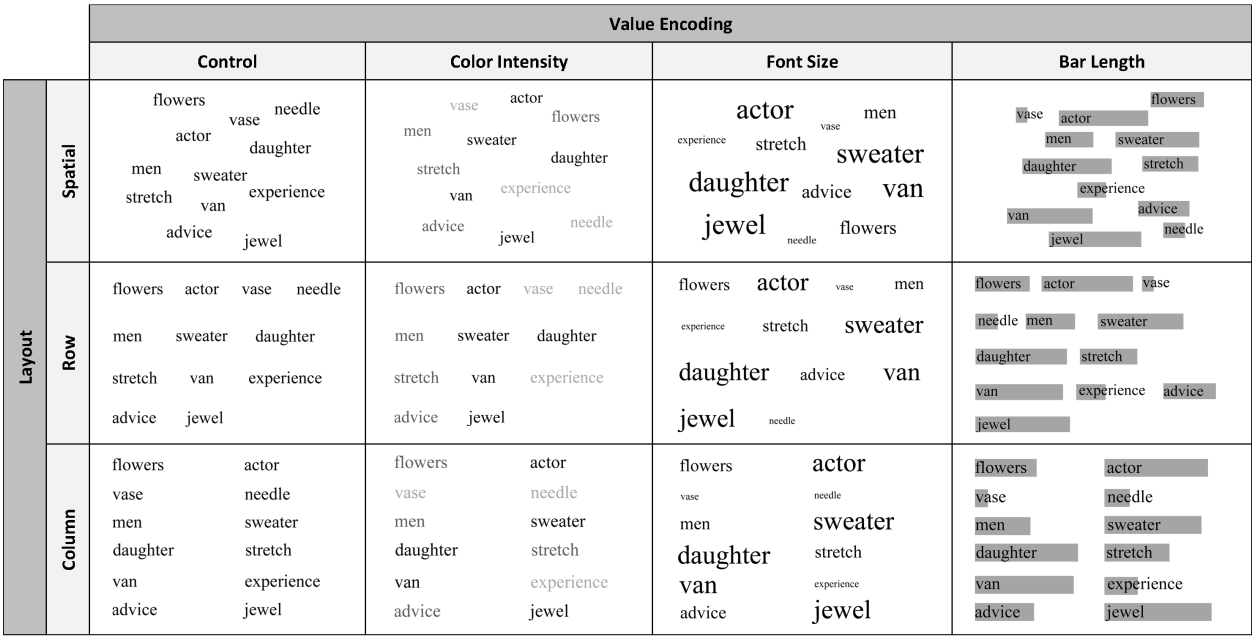


Figure 3.4: A visual summary of the word cloud formats made from the intersection of four types of visual encoding and three layouts. (This figure only serves as an example and the actual study trials contained more words than what is depicted here.)

Layout

We used three different layouts: *Spatial*, *Row*, and *Column*. The *spatial* layout positions the words randomly in the word cloud, the *row* layout aligns the words horizontally, and the *column* layout aligns the words vertically. The inclusion of these layouts in our study was inspired by the study by Felix et al. [14] who found that some layouts led to better performance compared to others in different tasks including magnitude judgment. By using similar layouts we hoped to see if the difference in performance was somehow related to certain user characteristics.

Value Encoding

The types of value encoding used in the word clouds are *color intensity*, *font size*, and *bar length* (shown in the columns of Figure 3.4). *Control* is also listed along the types of value encoding to show what the word clouds would look like without any specific form of visual encoding channel.

Graphical Marks: Control, color intensity, and font size all use the text itself as the graphical mark that represents the textual data, while the bar length uses the bar as the graphical mark and the text as a mere label for the bar. This differentiation allows us to test whether "additional" marks add value to interpreting word clouds for any of the user characteristics we test.

Visual Encoding Channels: The visual encoding channels that are used in our study are lightness of color in *color intensity*, the size of the words in *font size* and the length of the bar in *bar length* (shown in the rows of Figure 3.4). To test additional marks, other shapes could also have been used with attributes (channels) that suit them. For example, if a circle was used as a mark, the area of the circle would serve as the visual encoding channel. For this thesis, we have opted to use only bars, as according to previous research, bars have proven to be easier to interpret in comparison to circles [14].

Note that visual encoding and value encoding are used interchangeably in the thesis.

3.4 Implementation of the Experimental System

We implemented our experimental system on a website developed using MySQL as the database, and PHP as REST APIs for the communication between the front-end and back-end via stored procedures. The front-end accessed those APIs via Ajax calls and populated results in HTML5 via JavaScript. The website link was then posted to MTurk for users to access. Once the users went through the study procedure on the website, the last page of the website provided the users with a unique ID to enter on MTurk to signal to us that they have successfully completed the whole study. If for some reason the user was not able to complete the study due to connection issues, the website also had the ability to help the user continue from where they left off.

To make the word clouds we used the word cloud JavaScript class using the D3.js library written by Jason Davies [12]. We implemented multiple functions with different parameters to help make the different variations of word clouds.

The dataset we used was from the paper *"Taking Word Clouds Apart: An Empirical Investigation of the Design Space for Keyword Summaries"* by Felix et al. [14]. It included a total of 313 words and the topics related to health care reviews, an email collection and a collection of surveys, with the frequencies based on exponential distributions making the proportion between any two words constant and limiting the proportions to just 10 values [14].

The data with words and their frequencies was uploaded using a .csv file into the D3.js code. It was then visualized using Scalable Vector Graphics (SVG), with our functions allowing us to manipulate the word clouds based on the layout or visual encoding methods we chose. We were able to customize the SVG size, margins, padding, font sizes, font types, font colors, angles and more using these functions. We chose to make the word clouds on a canvas size of 500 by 400 pixels (See Appendix A for sample word clouds).

We wrote formulas to convert the frequency of the words to *font size*, *color intensity* and *bar length*. For *font size* we chose a minimum and maximum font size (which was entered as the parameters of the functions we wrote), then matched the frequencies of the words with the range of font sizes chosen, with the smallest font size corresponding with the least frequency and the largest font size corresponding with the highest frequency. Choosing a minimum font size allowed us to ensure that the smallest words are still legible for the participants.

For *color intensity* we chose 5 different shades of gray and divided the frequencies of the words into 5 different categories, matching words with higher frequencies to the darker shades of gray. For *bar length* we used a process similar to that of *font size*. We chose a minimum and maximum *bar length* (also entered as the parameters of the function) and then aligned the length of the bar with the frequency. Choosing a maximum bar length allowed us to ensure that our bars are not too long or falling off the SVG. We also ensured that no bars are overlapping each other so that the participant is not confused.

3.5 Method

3.5.1 Study Design

The study involved 9 word cloud designs based upon two independent variables: Layout types (3 layouts), and Visual Encoding methods (3 methods). Each design combination was repeated 4 times, resulting in a total of 36 trials (3 x 3 x 4). Table 3.1 gives an overview of the study variables and trials.

Task	Magnitude Judgement
Number of Participants	80-100
Design	Within Subject
User Characteristics	Perceptual Speed Verbal Working Memory
Design Elements	Layout = Row, Column, Spatial Visual Encoding = Color intensity, Font size, Bar length
Number of Trials	4 Word clouds x 3 layout types x 3 visual encoding methods = 36

Table 3.1: Overview of Study

The study had a within-subject design to ensure that every participant can be compared to each combination of *layout* type and *visual encoding* method. If we were to choose a between-subject design for this study, it would have been difficult to base our conclusions on the effect of a user's characteristics on the different combinations of word cloud designs.

Design elements (layout types and visual encoding methods) and user characteristics (perceptual speed and verbal working memory) were the independent variables of this study, while speed and accuracy were the dependent variables.

3.5.2 Participants

Due to the timing of this thesis, the study was conducted 100% online to ensure that all COVID-19 related regulations are followed. The use of MTurk allowed us to recruit a high number of participants in a small amount of time, which was quite beneficial at a time when the number of people that can gather in one place was limited.

A total of 84 people were recruited via Amazon Mechanical Turk (MTurk). Three of the participants were later found to have completed the study incorrectly, hence their data was not used as part of the analysis. This left a total of 81 participants for our data analysis.

MTurk has proven to be effective for testing visualizations [22] and more specifically word clouds [14]. It is a tool used by many studies due to its easy-to-use interface, security and degree of trust, verification of the identity of users, handling of payments, cost-effectiveness,

etc. Tests have been conducted to assess the quality of data retained from studies conducted on MTurk and have concluded that high-reputation workers rarely failed ACQs (attention check questions) and produced the highest-quality data [34].

The platform allowed us to recruit participants with a certain degree of reputation. To ensure response quality, the study was restricted to participants who have completed at least 100 HITs (Human Intelligence Tasks), received an approval rating of 95%, and received a Masters Qualification by MTurk. The study did not collect personally identifiable information and collected age, gender, and occupational information for analysis purposes.

3.5.3 Procedure

To start the study on MTurk, a consent form (See Appendix B) appeared to the participants. After providing their consent, participants were asked to complete a pre-study questionnaire (See Appendix C), where they provided their age, gender, occupational information, and their level of expertise with data visualizations.

The participants were then asked to complete standard tests of psychology to measure their *perceptual speed* and *verbal working memory*.

After completing the tests for measuring user characteristics, participants went through the task of *magnitude judgment* by judging the difference in frequency of words in a word cloud (See Figure 3.3). The Magnitude Judgement task consisted of 36 trials, and each trial

involved one word cloud and a corresponding question next to it. The questions were multiple choice containing 4 answers, one of which could be selected using the radio buttons, and when selected the participant would automatically be taken to the next page. The participant had a time limit of 15 seconds on each page.

Once all trials were completed, the participants were instructed to complete a short post-study questionnaire (See Appendix D) to provide their feedback regarding the word cloud designs and their overall experience throughout the study. In particular, we asked the participant to answer the following three questions about the visual encoding methods on a 5-point Likert scale: (1) I found bars (that depict the frequency of the words) useful when performing tasks, (2) I found the size of the words useful when performing tasks, and (3) I found the color saturation of words useful when performing tasks. We also asked them to rank these three visual encoding methods (bar, color, size) in terms of how difficult it was to judge the magnitude. The study procedure took approximately 40 minutes for each participant to complete. The participants were compensated after completing the study with the payments facilitated through MTurk.

3.6 Data Collection

As discussed before, during the study our experimental system collected the responses to questionnaires before and after the main study. The questionnaires included several closed-ended questions. During the main study, the scores for the perceptual speed and verbal

working memory tests were collected as quantitative data and stored in the database. Finally, while participants completed the trials with word clouds, we collected both the errors and task completion times. In the next chapter, we provide the analysis of the data we collected in detail.

Analysis and Results

In this section, we present the analysis and results on the data collected from the user study which help us to answer our two main research questions of *What to adapt?* and *How to adapt?* To analyze the data we consider three main measures, *log-error*, *time (s)* and *task performance*. For each of these measures, we analyzed how they get impacted by user characteristics, three layout types, and three visual encoding methods. We also look at a summary of the participants' background and an analysis of subjective measures. All data analyses were performed using IBM SPSS Statistics [15].

4.1 Participants' Background

4.1.1 Results of Pre-Study Questionnaire

Our study started with a series of pre-study questions to help us gather the background and demographic information of the participants. We asked the participants questions regarding their age, gender, occupation and how often they look at word clouds.

The participants of our study consisted of 43 males and 37 females with 1 person who chose not to disclose. Out of the five age groups listed, the age group with the most amount of participants (33.33%) was ages 35 to 44, while the age group with the least amount of participants (1.23%) was ages 18 to 24. Figure 4.1 summarizes the age group distribution.

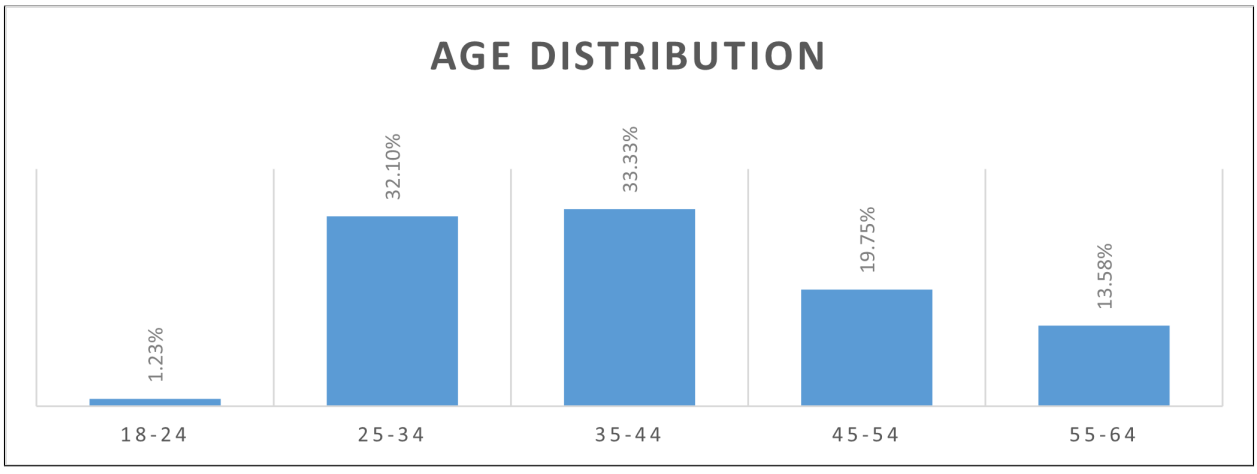


Figure 4.1: A bar graph summarizing the age distribution of participants in our study

We asked the participants how often they look at simple or complex word clouds, with

sample images showing them examples of simple and complex word clouds. Figure 4.2 summarizes the results. We noticed that while most participants encounter word clouds, they tend to look at simple word clouds more often than complex word clouds.

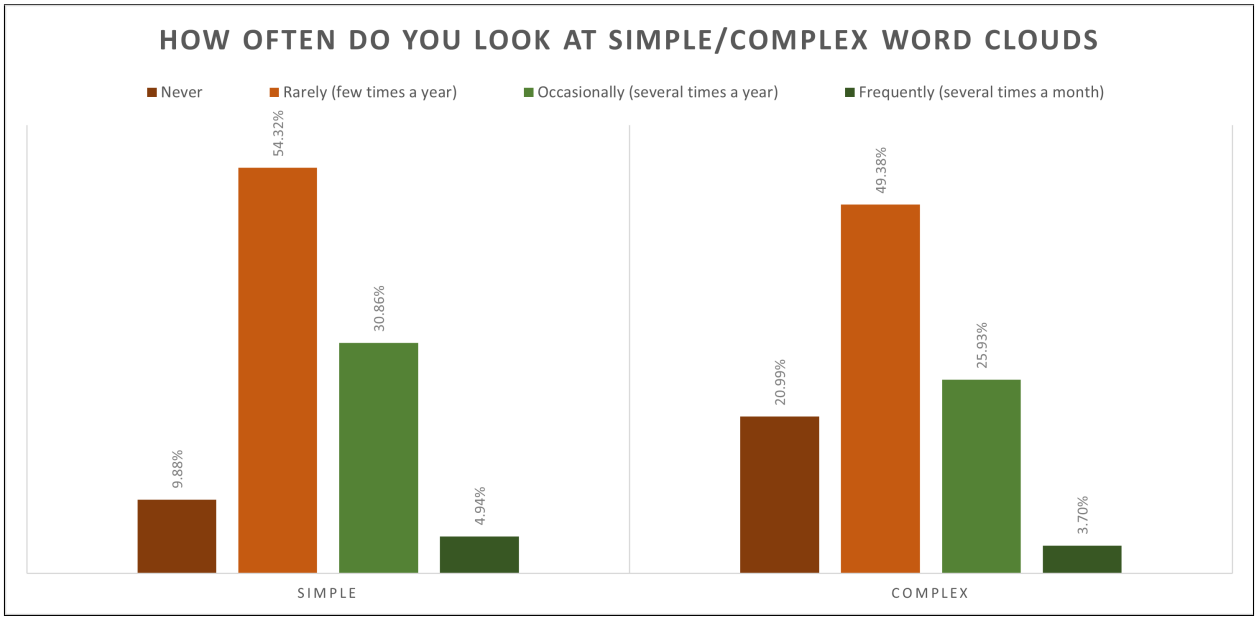


Figure 4.2: A summary of the results for the pre-study questions of *How often do you use simple bar graphs?* and *How often do you look at complex bar graphs?*

In terms of occupational background, we found that the participants came from a variety of backgrounds including teaching, accounting, software development, IT, data, project management, marketing, human resources, retail, customer service, etc. Out of all participants, about 9% identified themselves as students.

4.1.2 Results of User Characteristics Tests

Variable	Minimum	Maximum	Mean	Std. Deviation	Variance
Perceptual Speed	18	96	63.48	13.89	193.06
Verbal Working Memory	3	6	5.37	0.73	0.53

Table 4.1: Descriptive Statistics for Perceptual Speed and Verbal Working Memory

We tested the participants of our study for two user characteristics, *perceptual speed* and *verbal working memory*. The results for *perceptual speed* (M=63.48, SD=13.89) ranged between 18 and 96 (with 96 the maximum score possible) and the results for *verbal working memory* (M=5.37, SD=0.73) ranged between 3 and 6 (with 6 being the maximum score possible). Table 4.1 summarizes the descriptive statistics for *perceptual speed* and *verbal working memory*.

To simplify the analysis, we split the measures for the *perceptual speed* and *verbal working memory* test into three levels: *Low* represents the values within the bottom quartile (i.e., lower 25%), *Average* represents the values within the interquartile range (i.e., middle 50%), and *High* represents the upper quartile (top 25%).

4.2 Analysis of Log-Error and Time (s)

Variable	Minimum	Maximum	Mean	Std. Deviation	Variance
Time (s)	1	14	4.02	2.00	3.98
Log-Error	-3.00	5.29	1.05	3.05	9.31

Table 4.2: Descriptive Statistics for Time (s) and Log-Error

The data collected from the study is comprised of two main performance measures, i.e., speed and accuracy. The performance measure related to speed is the *time* (the time it took to select the answer in each of the 36 study trials, measured in seconds). The distribution of *time (s)* was normal (M=4.00, SD=2.00). The performance measure related to accuracy is *log-error* (normally distributed, M=1.05, SD=3.05). Table 4.2 summarizes lists the descriptive statistics for *time (s)* and *log-error*.

To understand *log-error*, we will first discuss *absolute error*. In the study, participants are asked to select 1 of 4 percentages listed to estimate the difference in the frequencies of the two highlighted words. The absolute error is a calculation of how far the participants were from the correct percentage. To calculate the absolute error, we used the following formula:
$error = |percentage_{true} - percentage_{estimated}|$

To calculate *log-error*, we then use the absolute error while adding 1/8 to avoid calculation errors when the *error* is 0. The formula used is:

$log\text{-}error = \log_2(error + 1/8)$.

Both the formulas for *error* and *log-error* have been used following the work from Felix et al. [14] on measuring performance on similar tasks with word clouds.

We use two separate Univariate General Linear Models [29] repeated measures for our two dependent variables: *log-error* and *time (s)*. This test was conducted to investigate the effects of our independent variables. Two univariate tests were chosen rather than one multivariate test due to the negative weak and statistically insignificant correlation of the two dependent variables ($r(2916)$=-.03, p=.104) based on a Pearson's correlation (r) test [16].

Effect sizes (partial eta squared) for all sections are reported as small for .01, medium for .06, and large as .14. For post hoc analysis, we use Tukey's Honest Significant Difference test to further compare the means after our initial significance tests [6]. We also use the Pearson's Chi-Square test in our analysis of subjective measures to evaluate categorical data [33].

4.2.1 Results of Log-Error

	Effect	**F-Ratio**	**Effect Size**	**Sig. Value**
Design Elements	Layout	F(2,240) = 13.522	η_p^2 = .101	p < .001
	Visual Encoding	F(2,240) = 2.389	η_p^2 = .020	p = .094
	Layout * Visual Encoding	F(4,720) = 18.292	η_p^2 = .023	p < .001
User Characteristics	Perceptual Speed	F(2,72) = 3.260	η_p^2 = 083	p < .05
	Verbal Working Memory	F(2,72) = .483	η_p^2 = .013	p = .619
	Perceptual Speed * Verbal Working Memory	F(4,72) = 1.526	η_p^2 = .078	p = .204

Table 4.3: Main and Interaction Effects for Log Error

For this analysis, we utilize a Univariate General Linear Model [29] using *log-error* as our dependent variable. Table 4.3 summarizes the results. From this table, we observe that there were significant main effects for *layout* and *perceptual speed* as well as significant interaction effects of *layout*visual encoding*. *Visual encoding* was found to be marginally significant. Overall, the results suggest that *layout* and *perceptual speed* significantly impact the performance of users in terms of errors they make.

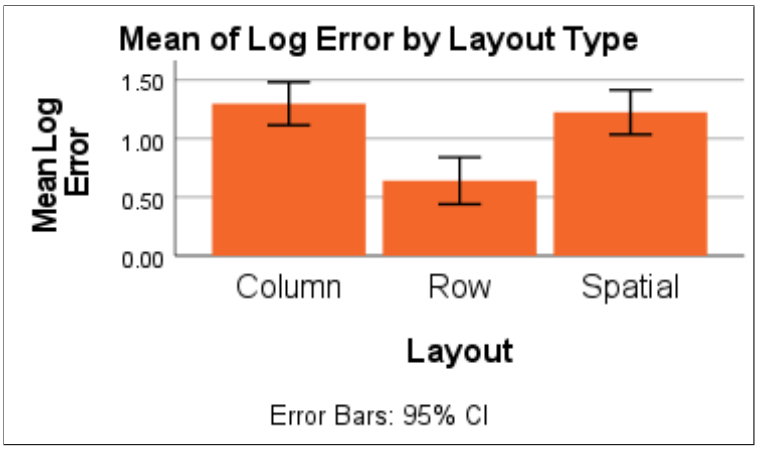


Figure 4.3: Mean Log-Error by Layout Type

A post hoc Tukey's HSD test confirmed the significant main effect of *layout* and showed that mean *log-error* is significantly lower ($p<.001$) when the *layout* is *row* (M=.64 , SD= 3.18) in contrast to both *spatial* (M=.1.22, SD= 3.01) and *column* (M=1.30 , SD= 2.91). The test also showed that the *spatial* and *column* groups did not differ significantly from each other. Figure 4.3 shows the difference in mean *log-error* for the different *layout* types. This suggests that the *row* layout negatively impacts the performance of users in terms of errors they make, however there is no significant difference between the other two layouts.

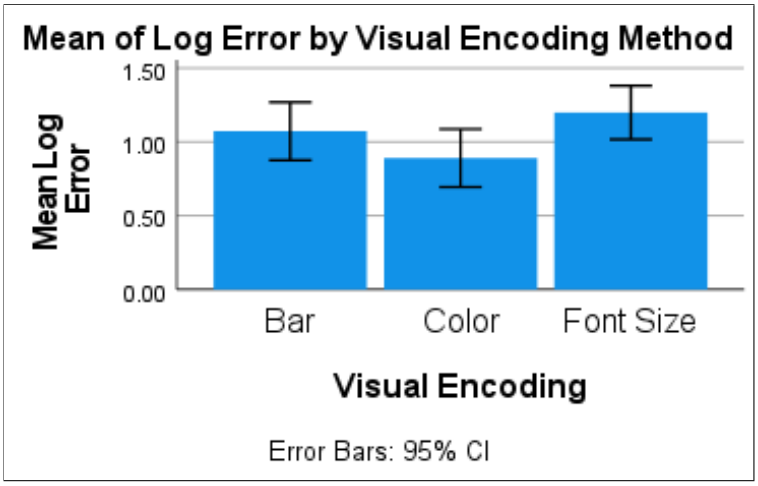


Figure 4.4: Mean Log-Error by Visual Encoding Method

While looking at the marginally significant effects of *visual encoding*, in figure 4.4 we observe that the use of *color intensity* (M=.89, SD=3.12) to encode word frequency gives the most accurate results on average, while using *font size* (M=1.2, SD=2.89) to encode frequency results in the highest amount of errors by users, and *bar length* (N=1.07, SD=3.13) remains in between.

The interaction effect of *layout* and *visual encoding* is highly visible when we look at *log-error* in contrast to all 9 combinations of the two variables. Figure 4.5 shows *row*color* having the least amount of errors (M=-.15, SD=3.19) while *spatial*font size* (M=1.70, SD=2.59) had the highest amount of errors. These results suggest that the usefulness of *visual encoding* methods generally depends on the type of layout that was used, which should be considered by the designers of word clouds.

As our study, of word clouds using the magnitude judgment task, is similar (in design)

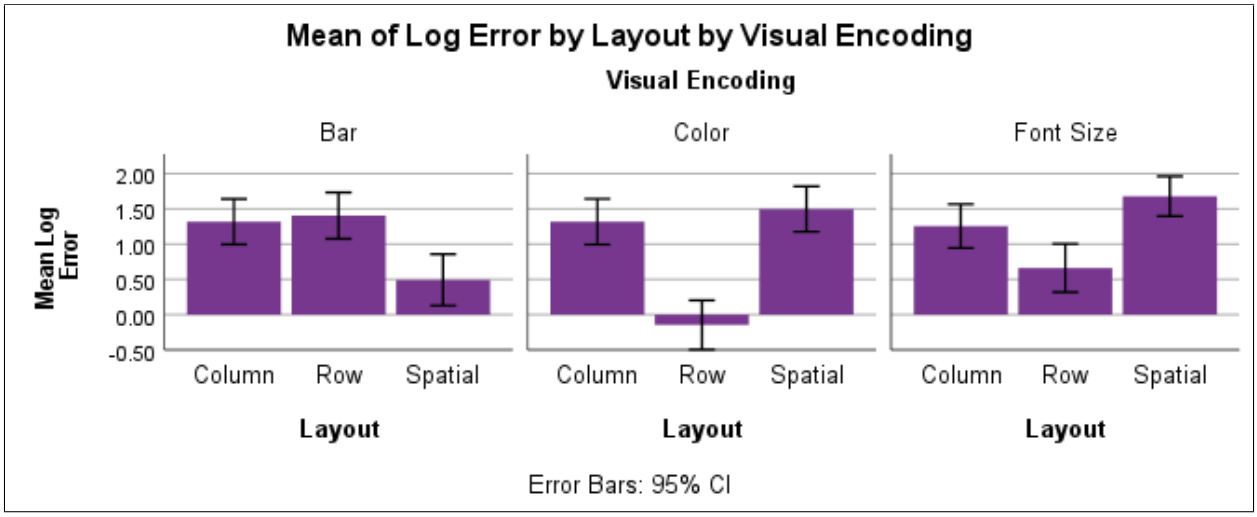


Figure 4.5: Mean Log-Error for Visual Encoding Methods by Layout Type

to the study by Felix et al. [14], we revisit their results and compare our findings with their findings. In particular, Felix et al. [14] found that the *spatial* layout impacted the performance of users negatively and that using additional marks (such as bars to encode word frequency) improved accuracy. Our study's results also aligned with these previous findings. When looking at *layout* types and *visual encoding* methods with accuracy as our measure, we see that the *spatial* layout was linked to the highest amount of errors when using *font size* and *color intensity*, while having the least amount of errors when using *bar length* as the *visual encoding* method.

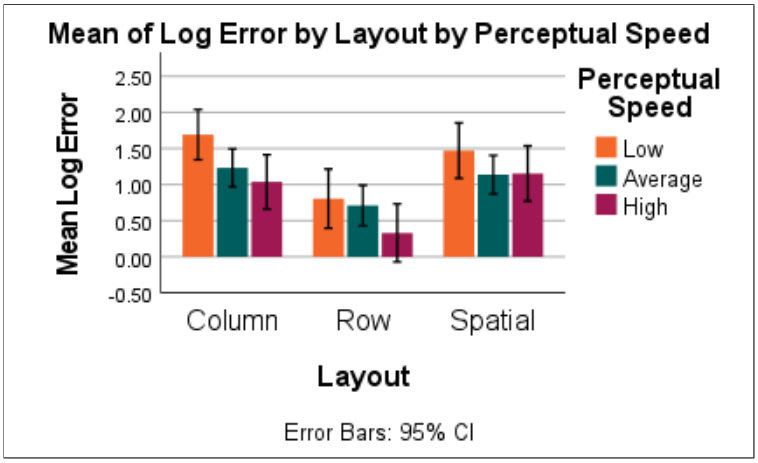


Figure 4.6: Mean Log-Error by Perceptual Speed for Layout

To analyze the significant main effects of *perceptual speed*, we look at how *perceptual speed* varies between *layout* types and *visual encoding* methods.

When comparing layout types in Figure 4.6 we see that those with *low perceptual speed* (M=.81, SD=3.22) can be seen performing almost as well as those with *average perceptual speed* (M=.71, SD=3.17) when we use the *row layout* type. Overall, we find that while *perceptual speed* usually impacts the performance of users across all layout types, the difference could be more pronounced when a certain layout is used.

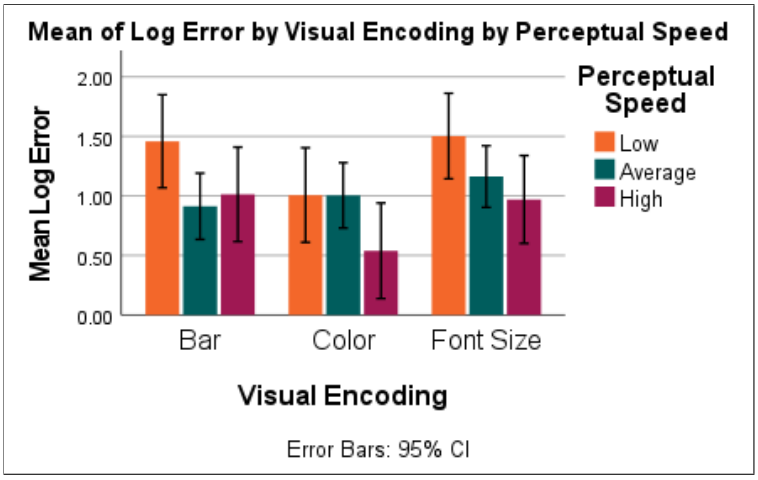


Figure 4.7: Mean Log-Error by Perceptual Speed for Visual Encoding

When comparing *visual encoding* methods in Figure 4.7 we see that when using *color intensity*, those with *low perceptual speed* (M=1.01, SD=3.12) performed just as well as those with *average perceptual speed* (M=1.01, SD=3.10). We also observe that those with *average perceptual speed* performed best when using the *bar length* (M=0.91, SD=3.14) method, when compared to *color intensity* (M=1.01, SD=3.10) and *font size* (M=1.16, SD=2.92). However this finding was not statistically significant.

Another non-significant observation is that, for participants with *high perceptual speed*, the *spatial layout* and the *bar length visual encoding* method proved to be the worst, as in both cases the graphs in Figures 4.6 and 4.7 show the group performing worse than the *average perceptual speed* group.

4.2.2 Results of Time (s)

	Effect	F-Ratio	Effect Size	Sig. Value
Design Elements	Layout	F(2,240) = .277	η_p^2 = .002	p = .758
	Visual Encoding	F(2,240) = 3.792	η_p^2 = .031	p < .05
	Layout * Visual Encoding	F(4,720) = .204	η_p^2 = .001	p = .936
User Characteristics	Perceptual Speed	F(2,72) = 7.312	η_p^2 = .169	p < .001
	Verbal Working Memory	F(2,72) = .584	η_p^2 = .016	p = .560
	Perceptual Speed * Verbal Working Memory	F(4,72) = 1.459	η_p^2 = .075	p = .224

Table 4.4: Main and Interaction Effects for Time (s)

We ran a Univariate General Linear Model [29] for *time (s)* as well. We found significant main effects of *visual encoding* method and *perceptual speed*. Recall that *perceptual speed* was also found to be significantly impacting the number of errors people make. However, unlike *layout* which was found to be impacting the number of errors people make, *visual encoding* influenced the participants in terms of task completion time.

A post hoc Tukey's HSD test ($p<.001$) confirmed the significant main effect of *visual encoding* methods when the dependent variable is *time (s)*. It showed that it takes the participants more time to judge the magnitude when using *bar length* (M=4.38 , SD= 2.184) in contrast to both *color intensity* (M=3.82, SD= 1.86) and *font size* (M=3.86, SD=1.87). Figure 4.8 shows this difference along with a breakdown of the *visual encoding* methods by the different levels of *perceptual speed*.

Another post hoc Tukey's test ($p<.001$) confirmed the significant main effect of *perceptual*

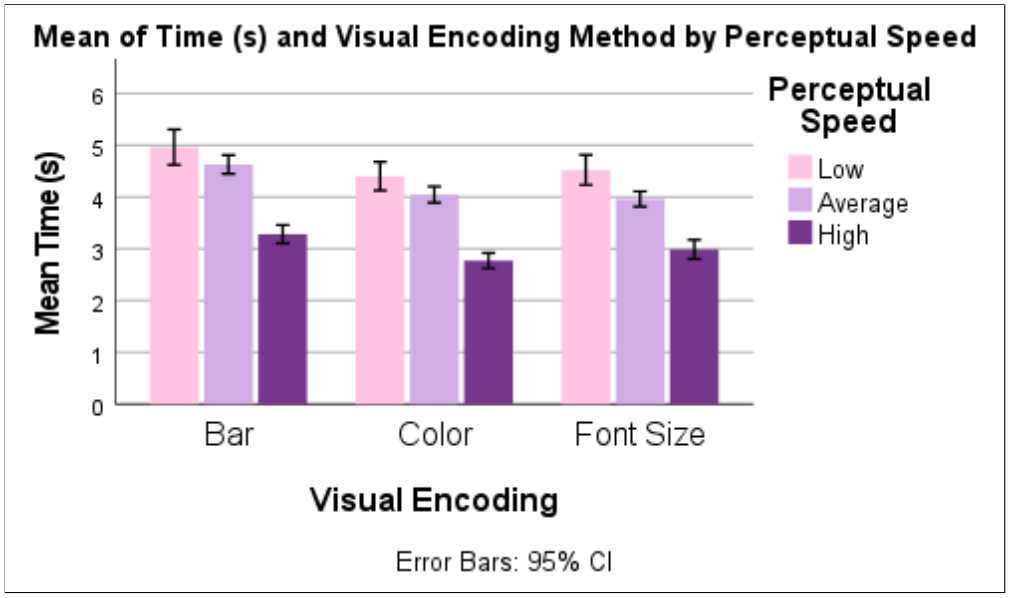


Figure 4.8: Mean Time (s) for Visual Encoding Method by Perceptual Speed

speed and *time (s)*, showing that all three levels of *perceptual speed* are significantly different from one another.

If we look at how the different levels of *perceptual speed* interact with the *visual encoding methods*, we see that in all three methods the levels (low, average, high) seem to maintain a similar ratio among themselves. There appears to only be a minor non-significant difference (based on Tukey's HSD) in the group with *high perceptual speed* performing slightly faster when the method is *color intensity* (M=2.77, SD=1.17) as compared to *font size* (M=2.99, SD=1.46). While the *average perceptual speed* group performs slightly faster when using the *font size* (M=3.96, SD=1.67) method as compared to *color intensity* (M=4.05, SD=1.77).

4.3 Analysis of Task Performance

The above-mentioned measures of *log-error* and *time (s)* are both useful when analyzing them individually, however, we also needed a way to analyze them together. For instance, taking less time to answer a question is better but it does not account for the accuracy of the participant. Similarly, correctly guessing the answer is good, but it does not show us how long it took the participant to guess the correct answer.

A way to tackle this is to create a measure that incorporates both speed and accuracy. In the paper by Carenini et al. [8], the two performance measures of *completion time* and *accuracy* have been combined to create a new performance score called *task performance*. This has been done to analyze *time (s)* but only after having penalized it for the incorrect trials. The formula is:

$$\textit{task performance (s)} = \frac{\text{Average Time for Correct Trial Repetitions}}{\text{Percentage of Correct Repetitions}}$$

The formula uses the average time for only the correct repetitions and adjusts it based on the percentage of correct repetitions. Each combination is repeated four times and the more incorrect repetitions there are in a combination (of *layout* and *visual encoding*), the higher the adjusted average time will be for that specific combination.

4.3.1 Results of Task Performance

	Effect	F-Ratio	Effect Size	Sig. Value
Design Elements	Layout	$F(2,225) = 2.545$	$\eta_p^2 = .022$	$p = .081$
	Visual Encoding	$F(2,227) = 2.423$	$\eta_p^2 = .021$	$p = .091$
	Layout * Visual Encoding	$F(8,598) = 4.058$	$\eta_p^2 = .051$	$p < .001$
User Characteristics	Perceptual Speed	$F(2,72) = 6.519$	$\eta_p^2 = .153$	$p < .05$
	Verbal Working Memory	$F(2,72) = .138$	$\eta_p^2 = .004$	$p = .871$
	Perceptual Speed * Verbal Working Memory	$F(4,72) = 1.819$	$\eta_p^2 = .092$	$p = .135$

Table 4.5: Main and Interaction Effects for Task Performance

We found that *Task performance* of participants was normally distributed (M=11.71, SD=7.39). Just like the previous sections, we also used a Univariate General Linear Model [29] for analyzing the *task performance*. When using *task performance* as the dependent variable, we found significant main effects of *perceptual speed*, and significant interaction effects of *layout* and *visual encoding*. *Layout* and *visual encoding* had marginally significant main effects. The effects are summarized in Table 4.5.

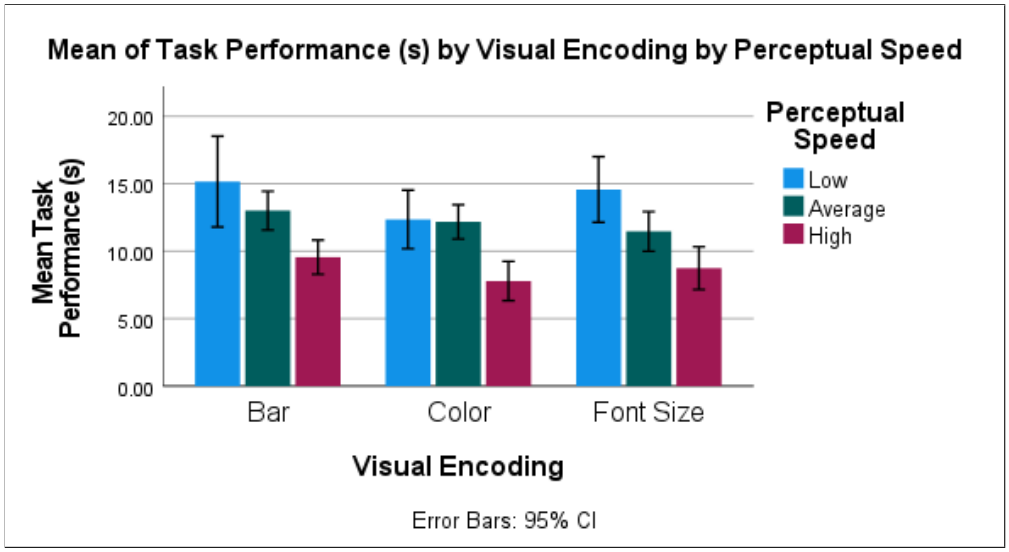


Figure 4.9: Mean Task Performance by Perceptual Speed for Visual Encoding Method

We first take a look at *perceptual speed*. When we analyze how user groups with different *perceptual speed* get impacted by different *visual encoding* methods (Figure 4.9, we see that all *perceptual speed* groups tend to perform better when the *color intensity* (M=11.10, SD=6.89) is used as a *visual encoding* method. Note that within the *color intensity* category, those with *low perceptual speed* (M=12.36, SD=7.56) perform almost as well as those with *average perceptual speed* (M=12.18, SD=6.77).

We also notice that those with *average perceptual speed* do perform slightly better in the *font size* category (M=11.47, SD=7.09) when compared to the *color intensity* category (M=12.8, SD=6.77). However, this finding was not statistically significant based on a Tukey's HSD test.

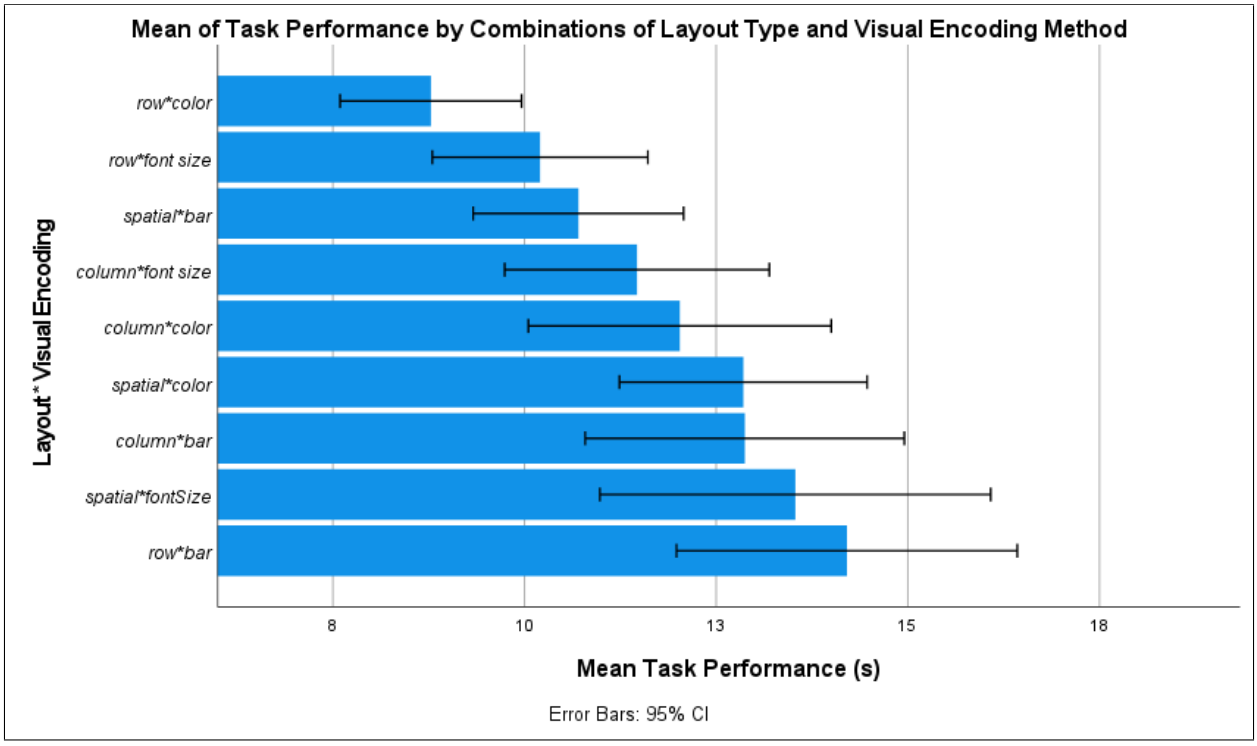


Figure 4.10: Mean Task Performance by Combinations of Layout Type and Visual Encoding Method

We found significant interaction effects for *layout* and *visual encoding*. This time we look at *task performance* and the nine combinations of *layout* types and *visual encoding* methods. As can be seen in Figure 4.10, overall the *row*color* combination had the best mean *task performance (s)* (M=8.78, SD=5.32), while the *row*bar* category had the worst mean *task performance (s)* (M=14.21, SD=9.04). A post hoc Tukey's HSD test showed *row*bar* was significantly different from both *row*color* ($p<.001$) and *row*font size* size ($p<.05$). However, the difference between *row*color* and *row*font size* was not significant.

4.4 Analysis of Subjective Measures

4.4.1 Ranking Visual Encoding Methods

In the post-study questionnaire, we asked the participants to rank the three types of *visual encoding* methods used from their most preferred to their least preferred (See Appendix C). Overall participants ranked *font size* (45.7%) as the most preferred, followed by *color intensity*, and then *bar length* as their least preferred *visual encoding* method.

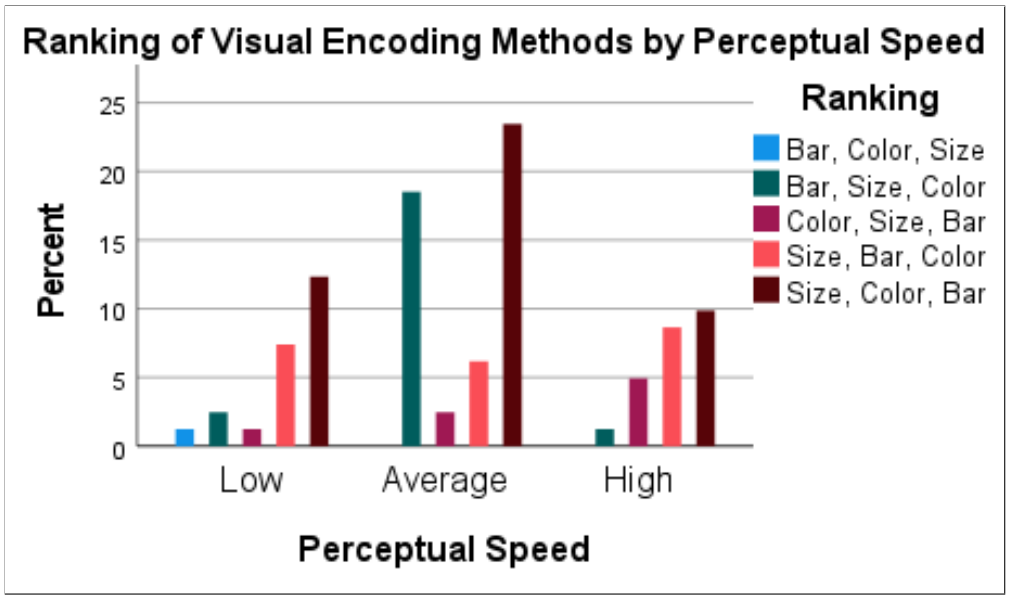


Figure 4.11: Ranking of Visual Encoding Methods by Perceptual Speed

We ran a Pearson Chi-Square between our user characteristics and the ranking preferences and found a significant relationship between *perceptual speed* and the rankings ($X^2(8, N=81)=18.93$, p=.<05). Previous literature [8, 42] also found significant relationships

when comparing user characteristics and user preferences on different types of visualizations, however, this is the first time such findings have been confirmed for word clouds.

Figure 4.11 shows those with *average perceptual speed* preferred *bar length* (36.6%) before *font size* and *color intensity*. Just like *high perceptual speed* and *low perceptual speed*, most (46.3%) still preferred *font size* followed by *color intensity* and *bar length*. These results align with the results of our measures, with *font size* performing the best, and *bar length* performing the worst.

It should be noted that participants were given 6 rankings to choose from, however, no one chose the *color, bar, size* order of ranking.

4.4.2 Usefulness Ratings of Visual Encoding Methods

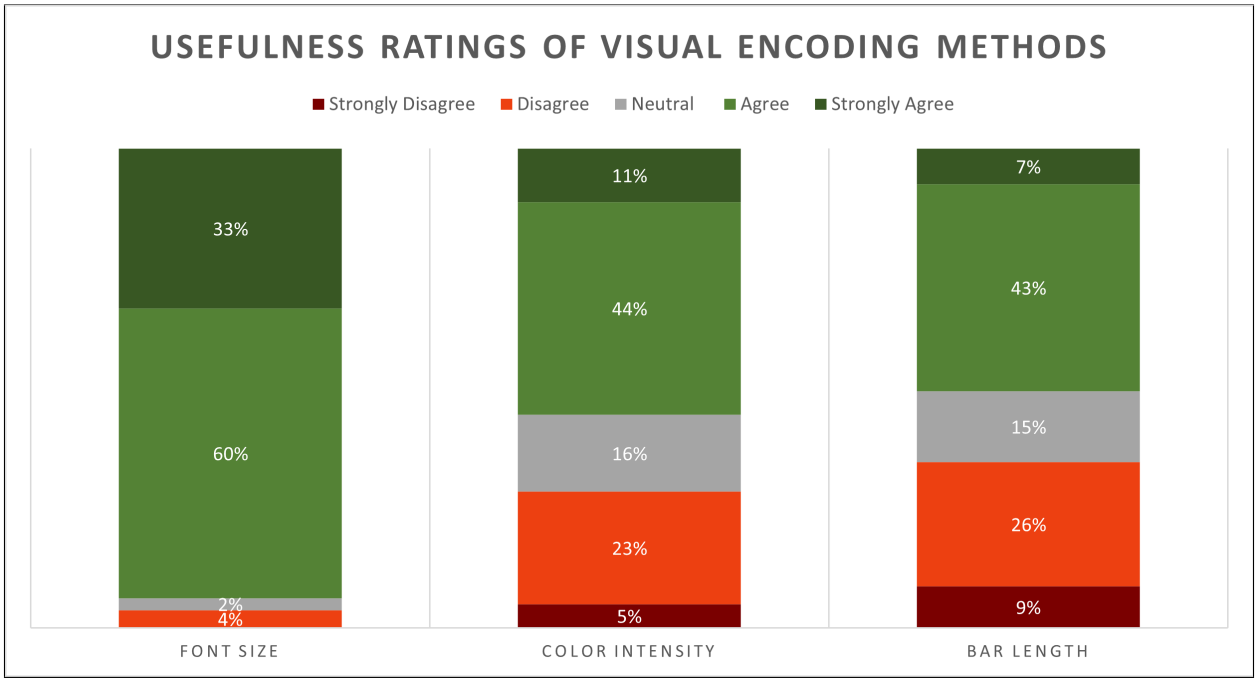


Figure 4.12: Stacked bars summarizing the usefulness ratings of Visual Encoding Methods

We asked the participants to share if they found the three *visual encoding* methods useful. The answers were choices on a Likert scale ranging from "Strongly Agree" to "Strongly Disagree" (See Appendix C). Figure 4.12 presents a summary of the responses we collected for each of the three types of *visual encoding* methods.

The results showed that 93.8% of participants either agreed or strongly agreed that they found the size of the words useful when performing the tasks.

However, the results for the usefulness of *color intensity* and *bar length* were quite close to each other, with more people agreeing that they found *color intensity* useful when performing

the tasks.

4.5 Summary

In this chapter, we analyzed the data that we collected from the user study we ran on Amazon Mechanical Turk. For analysis, we used three different measures: *log-error*, *time (s)* and *task performance (s)*. Through various statistical analyses, we found that *perceptual speed* impacts the performance of users across various measures. We also observed that *perceptual speed* may impact users differently depending on the type of *layout* or *visual encoding* method that is used. Finally, we found that *perceptual speed* may also impact the way people rank their preference for *visual encoding* methods.

Conclusions and Future Work

In this chapter, we first make some concluding remarks including the summary of our findings as well the importance of these findings. We then provide possible future directions for this work.

5.1 Conclusions

The goal of this thesis was to explore the word cloud design space and to explore whether there are significant relationships between word cloud designs and user characteristics and if so, how user characteristics may impact user performance for different designs. We chose word clouds for our study because, according to our knowledge, while other forms of common visualizations (such as bar graphs) have been researched with an emphasis on understanding the impact of user characteristics, there is no prior research based on user differences when it comes to word clouds. Thus the findings from this thesis can be important for future research

in this area.

We now revisit our initial research questions. In our study, we wanted to find out (1) what user characteristics are impacted by different design variations of word clouds; and (2) how word clouds can be improved to support users with specific characteristics. We answered these questions by conducting a study involving the magnitude judgment task across nine different word cloud variations made from different layout types and visual encoding methods.

To answer our first question, our research involved the testing of two main user characteristics: *perceptual speed* and *verbal working memory*. Of these two, we found significant effects of *perceptual speed* in the *log-error* and *task performance (s)* measures.

Verbal working memory did not have any significant effects based on the tests conducted. This could be due to the task chosen for our study. The task of *magnitude judgment* does not involve recalling any of the words, rather it is a quick judgment of the difference of frequency of the two highlighted words (mainly by judging the color intensity, font size, or the length of the bar) and the focus of the user is the graphical marks and value encoding channels. Further research with more complex tasks is necessary to confirm whether verbal working memory may impact user performance while using word clouds.

To answer our second question we first discuss the effect of *perceptual speed* in more detail. When we analyzed the different levels *perceptual speed* (low, average, high) and the three types of *layout* (row, column, spatial), results showed that those with *low perceptual speed* performed just as well as those with *average perceptual speed* when using the *row* layout.

Similarly, when we analyzed *perceptual speed* alongside the three *visual encoding* methods (color intensity, font size, bar length), results showed that those with *low perceptual speed* performed just as well as those with *average perceptual speed* when using the *color intensity* method.

There was also a significant relationship between *perceptual speed* and rankings of preferred *visual encoding* methods. We found that those with *average perceptual speed* chose *bar length* then *font size*, and then *color intensity* as their preferred order, while those with *high* or *low perceptual speed* chose *font size* first, then *color intensity* and then *bar length* as their least preferred *visual encoding* method. For usefulness ratings, participants also ranked *font size* as the most useful, and *bar length* as the least useful.

Word cloud design elements presented significant results as well and would be key in improving word clouds to support users. Our results showed that participants performed significantly better with the *row* layout compared to the *spatial* and *column* layouts for all types of users based on the *log-error* measure.

Analysis of *visual encoding* methods showed that it took significantly more *time (s)* to use *bar length* to judge magnitude than the other two (*color intensity* and *font size*) methods. This could have been because in the case of *bar length* the bar is a graphical mark (contrary to color intensity and font size, where the word itself is the graphical mark), causing the user to take more time to analyze the word cloud. The length of the word could also have confused the user when trying to judge the length of the bar as the words were placed on top

of the bars.

The interaction of both *layout* types and *visual encoding* methods significantly showed *row*color* followed by *row*font* size as the combinations with the best *task performance (s)*. *Row*bar* resulted in the worst *task performance (s)*. These results also aligned with the results of the subjective measures. *Font size*, *color intensity* then *bar length* was the most preferred order selected by the participants in the post study questionnaire.

We believe that the above results have implications for user-adaptive visualizations, where we can focus on tailoring word cloud design with respect to individual needs and differences. For example, a start point could be to develop an interactive word cloud interface that allows users to change the *layout* or *visual encoding* methods of a word cloud. Another step could be to predict user characteristics automatically by collecting the interaction data of a user so that a version of the word cloud that better matches the user's characteristics is shown to the user.

5.2 Future Work

Since, to our knowledge, this thesis presents the first study on exploring the impact of user characteristics on word clouds, there are several key avenues for future work.

First, in this thesis, we investigated the performance of users using the magnitude judgment task. There are still several other tasks that can be designed for both social and analytical

uses of word clouds. For example, topic discovery, topic matching, and keyword search are some of the tasks that have been tested to understand the effectiveness of different layouts and visual encoding methods for word clouds [14]. However, these tasks have not been tested while considering user characteristics as factors. As a starting point, we focused on the low-level task of magnitude judgment, however, we anticipate that with more complex tasks such as topic matching and keyword search, the influence of different user characteristics could be even more pronounced.

Second, previous research investigating the impact of user characteristics on other visualizations such as bar charts [41] has considered a variety of user characteristics such as reading proficiency, visualization literacy, need for cognition, verbal IQ, visual working memory, spatial memory, etc. In order to keep the study session under a reasonable time limit, we considered two important user characteristics (perceptual speed and verbal working memory). However, in the future, it would be interesting to observe how other cognitive abilities, as well as personality traits, may impact the user's performance while interacting with word clouds.

Third, With respect to exploring visual encoding methods, we have experimented with font size, color intensity, and bar length. However, other visual encoding methods can be examined in the future such as the size of circles to encode the frequency of words in a word cloud [14]. Similarly, other forms of visual encoding methods, semantic groupings, and layouts can be experimented with respect to user characteristics.

Fourth, in terms of word cloud algorithms, it would be interesting to explore different

keyphrase extraction techniques. Since, unlike individual words, keyphrases may allow us to better understand the context of word usage, future work could explore how people with varying user characteristics would perform with phrase clouds compared to basic word clouds.

Fifth, the study was conducted on MTurk which has been considered to be a reliable platform for hiring participants [22, 34] and was used in other studies on word cloud visualizations, on similar tasks [14]. However, the data collected from MTurk *crowdworkers* may be prone to sample bias, for instance with the age distribution or expertise of users with word clouds. In the future we hope to run studies with a larger and more diverse range of users to improve the generalizability of our research to address any bias problems.

Last but not least, future work may examine how the influence of user characteristics on different word cloud designs can be adapted to applications. Future research could use machine learning techniques to analyze the user's interaction data (mouse interactions, eye gaze) and predict their characteristics. Some previous research has already demonstrated how eye gaze data can be used to infer visualization tasks and cognitive abilities [39]. Following such research, machine learning models can be employed to predict the user's cognitive abilities and then subsequently provide adaptive interventions, for example by adapting to a more suitable layout or visual encoding. We hope that further research in the area will help to make word clouds to be more accessible to those with specific needs or levels of expertise.

Bibliography

[1] *Abcya*. URL: https://www.abcya.com/games/word_clouds.

[2] Aretha B Alencar, Maria Cristina F de Oliveira, and Fernando V Paulovich. "Seeing beyond reading: a survey on visual text analytics". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2.6 (2012), pp. 476–492.

[3] *APA Dictionary of Psychology*. URL: https://dictionary.apa.org/need-for-cognition.

[4] *APA Dictionary of Psychology*. URL: https://dictionary.apa.org/locus-of-control.

[5] AD Baddeley. *Working Memory Clarendon Press*. 1986.

[6] Jeff Austin Brillinger. *The Collected Works of John W. Tukey: Time*. Vol. 1. CRC Press, 1984.

[7] John T Cacioppo, Richard E Petty, and Chuan Feng Kao. "The efficient assessment of need for cognition". In: *Journal of personality assessment* 48.3 (1984), pp. 306–307.

[8] Giuseppe Carenini et al. “Highlighting interventions and user differences: informing adaptive information visualization support”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2014, pp. 1835–1844.

[9] Jason Chuang, Christopher D. Manning, and Jeffrey Heer. ““Without the Clutter of Unimportant Words”: Descriptive Keyphrases for Text Visualization”. In: *ACM Trans. Comput.-Hum. Interact.* 19.3 (Oct. 2012). ISSN: 1073-0516. DOI: 10.1145/2362364.2362367. URL: https://doi.org/10.1145/2362364.2362367.

[10] Cristina Conati and Heather Maclaren. “Exploring the role of individual differences in information visualization”. In: *Proceedings of the working conference on Advanced visual interfaces*. 2008, pp. 199–206.

[11] Cristina Conati et al. “Evaluating the impact of user characteristics and different layouts on an interactive visualization for decision making”. In: *Computer Graphics Forum*. Vol. 33. 3. Wiley Online Library. 2014, pp. 371–380.

[12] Jason Davies. *jasondavies/d3-cloud*. 2018. URL: https://github.com/jasondavies/d3-cloud.

[13] Ruth B Ekstrom. *Kit of factor-referenced cognitive tests*. Educational Testing Service, 1976.

[14] Cristian Felix, Steven Franconeri, and Enrico Bertini. “Taking word clouds apart: An empirical investigation of the design space for keyword summaries”. In: *IEEE transactions on visualization and computer graphics* 24.1 (2017), pp. 657–666.

[15] Andy Field. *Discovering statistics using IBM SPSS statistics*. sage, 2013.

[16] David Freedman, Robert Pisani, and Roger Purves. "Statistics (international student edition)". In: *Pisani, R. Purves, 4th edn. WW Norton & Company, New York* (2007).

[17] Martin J Halvey and Mark T Keane. "An assessment of tag presentation techniques". In: *Proceedings of the 16th international conference on World Wide Web*. 2007, pp. 1313–1314.

[18] Jason Hanna. *Expert: Thanksgiving could be "mother" of superspreader events*. Nov. 2020. URL: `https://www.cnn.com/2020/11/26/health/us-coronavirus-thursday/index.html`.

[19] Marti Hearst. *Search user interfaces*. Cambridge university press, 2009.

[20] Marti Hearst et al. "An evaluation of semantically grouped word cloud designs". In: *IEEE Transactions on Visualization and Computer Graphics* (2019).

[21] Marti A Hearst and Daniela Rosner. "Tag clouds: Data analysis tool or social signaller?" In: *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008)*. IEEE. 2008, pp. 160–160.

[22] Jeffrey Heer and Michael Bostock. "Crowdsourcing graphical perception: using mechanical turk to assess visualization design". In: *Proceedings of the SIGCHI conference on human factors in computing systems*. 2010, pp. 203–212.

[23] Michael J Kane and Randall W Engle. "The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective". In: *Psychonomic bulletin & review* 9.4 (2002), pp. 637–671.

[24] Byron YL Kuo et al. "Tag clouds for summarizing web search results". In: *Proceedings of the 16th international conference on World Wide Web*. 2007, pp. 1203–1204.

[25] Sébastien Lallé, Dereck Toker, and Cristina Conati. "Gaze-Driven Adaptive Interventions for Magazine-Style Narrative Visualizations". In: *IEEE Transactions on Visualization and Computer Graphics* (2019).

[26] Decong Li et al. "A Semi-Supervised Key Phrase Extraction Approach: Learning from Title Phrases through a Document Semantic Network". In: *Proceedings of the ACL 2010 Conference Short Papers*. ACLShort '10. Uppsala, Sweden: Association for Computational Linguistics, 2010, pp. 296–300.

[27] Taiyu Lin. "Web-OSPAN: A Web-based Working Memory Task". In: ().

[28] Robert H Logie. "Spatial and visual working memory: A mental workspace". In: *Psychology of learning and motivation*. Vol. 42. Elsevier, 2003, pp. 37–78.

[29] Keith A McNeil, Isadore Newman, and Francis J Kelly. *Testing research hypotheses with the general linear model*. SIU Press, 1996.

[30] *MTurk FAQs*. URL: `https://www.mturk.com/worker/help`.

[31] Tamara Munzner. *Visualization analysis and design*. CRC press, 2014.

[32] Chau Q. Nguyen and Tuoi T. Phan. "An Ontology-Based Approach for Key Phrase Extraction". In: *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*. ACLShort '09. Suntec, Singapore: Association for Computational Linguistics, 2009, pp. 181–184.

[33] Karl Pearson. "X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 50.302 (July 1900), pp. 157–175. DOI: 10.1080/14786440009463897. URL: https://doi.org/10.1080/14786440009463897.

[34] Eyal Peer, Joachim Vosgerau, and Alessandro Acquisti. "Reputation as a sufficient condition for data quality on Amazon Mechanical Turk". In: *Behavior research methods* 46.4 (2014), pp. 1023–1031.

[35] Anna W Rivadeneira et al. "Getting our head in the clouds: toward evaluation studies of tagclouds". In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. 2007, pp. 995–998.

[36] Johann Schrammel. "Exploring new ways of utilizing automated clustering and machine learning techniques in information visualization". In: *IFIP Conference on Human-Computer Interaction*. Springer. 2011, pp. 394–397.

[37] Johann Schrammel, Michael Leitner, and Manfred Tscheligi. "Semantically structured tag clouds: an empirical evaluation of clustered presentation approaches". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2009, pp. 2037–2040.

[38] James Sinclair and Michael Cardew-Hall. "The folksonomy tag cloud: when is it useful?" In: *Journal of Information Science* 34.1 (2008), pp. 15–29.

[39] Ben Steichen, Giuseppe Carenini, and Cristina Conati. "User-adaptive information visualization: using eye gaze data to infer visualization tasks and user cognitive abilities". In: *Proceedings of the 2013 international conference on Intelligent user interfaces*. 2013, pp. 317–328.

[40] Daniel Steinbock. *TagCrowd*. URL: https://tagcrowd.com/.

[41] Dereck Toker. "User characteristics and eye tracking to inform the design of user-adaptive information visualizations". PhD thesis. University of British Columbia, 2019. DOI: http://dx.doi.org/10.14288/1.0380857. URL: https://open.library.ubc.ca/collections/ubctheses/24/items/1.0380857.

[42] Dereck Toker et al. "Individual user characteristics and information visualization: connecting the dots through eye tracking". In: *proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2013, pp. 295–304.

[43] Ursula Torres Parejo et al. "A survey of tag clouds as tools for information retrieval and content representation". In: *Information Visualization* 20.1 (2021), pp. 83–97.

[44] Neil Turner. *Say it with a word cloud*. June 2011. URL: http://www.uxforthemasses.com/word-clouds/#.

[45] Fernanda B Viegas, Martin Wattenberg, and Jonathan Feinberg. "Participatory visualization with wordle". In: *IEEE transactions on visualization and computer graphics* 15.6 (2009), pp. 1137–1144.

[46] Martin Wattenberg and Fernanda B Viégas. "The word tree, an interactive visual concordance". In: *IEEE transactions on visualization and computer graphics* 14.6 (2008), pp. 1221–1228.

[47] *What is perceptual speed.* URL: http://www.igi-global.com/dictionary/perceptual-speed/42515.

[48] *Word Cloud Generator*. 2020. URL: https://www.jasondavies.com/wordcloud/.

[49] *WordArt.* URL: https://https://wordart.com/.

[50] *WordCloud Generator by MonkeyLearn.* URL: https://monkeylearn.com/word-cloud/.

[51] *Wordclouds.com.* URL: https://https://wordclouds.com/.

[52] *WordItOut.* URL: https://https://worditout.com/word-cloud/create/.

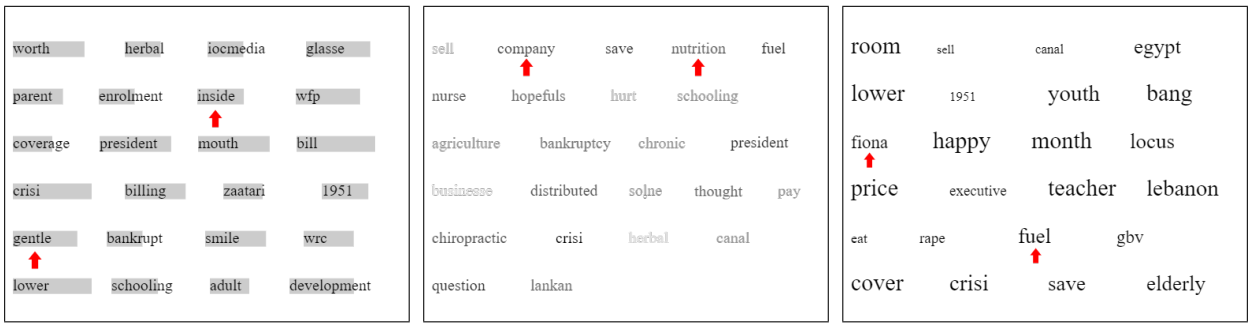
[53] Caroline Ziemkiewicz et al. "How locus of control influences compatibility with visualization style". In: *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE. 2011, pp. 81–90.

Appendices

Appendix A: Sample Word Clouds

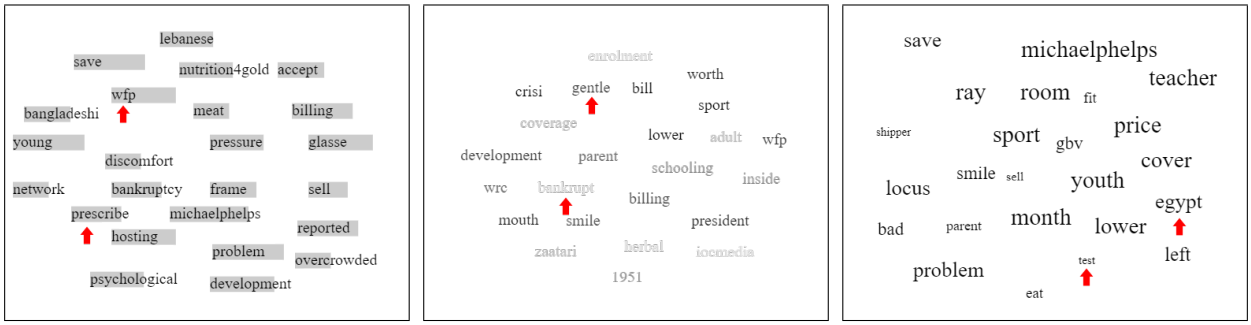
(a) Bar Length (b) Color Intensity (c) Font Size

Figure 1: Column Layout



(a) Bar Length (b) Color Intensity (c) Font Size

Figure 2: Row Layout



(a) Bar Length (b) Color Intensity (c) Font Size

Figure 3: Spatial Layout

Appendix B: Informed Consent Form

Date: _____

Name of Participant: _____

Study Name: Study of the effect of user characteristics on perceiving word clouds

Researcher: Zehra Shirin, Graduate Student in Information Systems and Technology, York University, zshirin@my.yorku.ca

Purpose of the Research: In this research, we are interested in studying the effects of user characteristics in perceiving word clouds. Humans have different cognitive abilities and personality traits. Previous research has shown that these differences affect the way a person perceives a visualization, specifically their ability to extract valuable information from the visualization through analytical or search based tasks. In this research, we are designing word clouds (a type of text visualization) using different layouts and methods and testing these word clouds against users with differing user characteristics. This study will help us to find valuable information on how word cloud designs could adapt to the needs of the user based on their abilities. To conduct this study, we will ask you to complete some tests of psychology and then complete a task on multiple word clouds. Then we will use the collected data to come to conclusions regarding whether user characteristics affect the way a person perceives a word cloud and how the word cloud could be improved to help people with certain user characteristics (For example, how a word cloud design could be adapted to the needs of users with low perceptual speed).

What You Will Be Asked to Do in the Research: In this research you will be asked to perform the following tasks: 1. At first, a consent form will appear to you before the experimental session begins. 2. After providing the consent, you will be asked to complete a pre-study background questionnaire about your educational background, age range and language proficiency. 3. You will then be asked to complete tests of psychology for perceptual speed, verbal working memory and visual working memory. 3. Each HIT will involve viewing a word cloud visualization and answering a question about it. 4. The study session will involve 36-48 HITs and will last between 40-50 minutes. You will receive $1.0 for completing each HIT. 5. After completing the HITs, you will be asked to fill a post-study questionnaire about the questions that you were asked during the study.

Risks and Discomforts: We do not foresee any risks or discomfort from your participation in the research.

Benefits of the Research and Benefits to You: Taking part in this research study may not benefit you personally, but we may learn new things that could help others.

Voluntary Participation and Withdrawal: Your participation in the study is completely voluntary and you may choose to stop participating at any time. If you stop participating at any point, you will still be eligible to receive the compensation for HITs you have already completed. In the event you withdraw from the study, all associated data collected will be immediately destroyed wherever possible.

Confidentiality: All personally identifying information you supply during the research – namely: your MTurk ID – will be held in confidence and will not appear in any report or publication of the research. Please be aware that your MTurk Worker ID can potentially be linked to information about you on your Amazon public profile page, depending on the settings you have for your Amazon profile. We will not be accessing any personally identifying information about you that you may have put on your Amazon public profile page. Your personally identifying data will be safely stored

in a locked facility and a password protected digital medium (USB key or DVD) and only research staff/research team members will have access to this information. The personally identifying data will be destroyed by the end of 2021. Confidentiality will be provided to the fullest extent possible by law.

Your non-personally identifying responses – including demographics (age-range, educational background and questionnaire responses – may be published on the PI's web-sites (personal, lab) and shared with the research community as data files. The PI will screen the data and use her own judgement to identify accidentally or indirectly personally identifying features and will redact the parts in question or, if necessary, the entire response prior to availing it to third parties. The anonymized non-personally identifying data collected in this research project may also be used by members of the research team in subsequent research investigations exploring similar lines of inquiry. The researcher(s) acknowledge that the host of the online survey (e.g., Amazon Mechanical Turk) may automatically collect participant data without their knowledge (i.e., IP addresses). Further, because this project employs e-based collection techniques, data may be subject to access by third parties as a result of various security legislation now in place in many countries and thus the confidentiality and privacy of data cannot be guaranteed during web-based transmission.

Questions About the Research? If you have questions about the research in general or about your role in the study, please feel free to contact Zehra Shirin by e-mail (zshirin@my.yorku.ca) or the Information Systems and Technology Graduate Program Office by phone (416-736-2100 ext. 40747) or email (lapsitec@yorku.ca). This research has received ethics review and approval by the Human Participants Review Sub-Committee, York University's Ethics Review Board and conforms to the standards of the Canadian Tri-Council Research Ethics guidelines. If you have any questions about this process, or about your rights as a participant in the study, please contact the Sr. Manager and Policy Advisor for the Office of Research Ethics, 5th Floor, Kaneff Tower, York University (telephone 416-736-5914 or e-mail ore@yorku.ca). **Consent:** By clicking "Agree" below, you confirm that you have read the consent form, are at least 18 years old, and agree to participate in the study "Study of the effect of user characteristics on perceiving word clouds".

☐ I agree to participate in the research

☐ I do NOT agree to participate in the research

Appendix C: Pre-Study Questionnaire

Please answer each of the following questions.

Please select your age range:

- ☐ 18-24
- ☐ 25-34
- ☐ 35-44
- ☐ 45-54
- ☐ 55-64

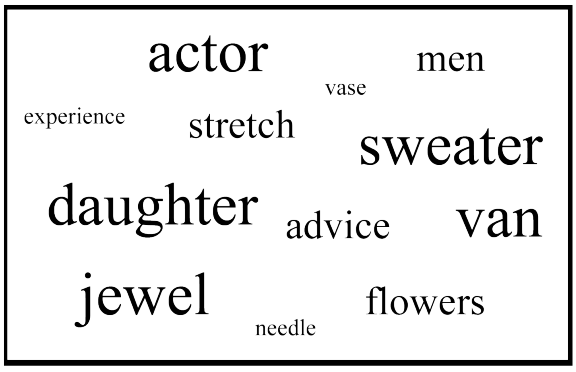
Please select your gender:

- ☐ Male
- ☐ Female
- ☐ Rather not say

Occupation: _____

Field of Study (if student, otherwise enter n/a): _____

How often do you look at simple Word Clouds, like the one below:



- ☐ Never
- ☐ Rarely (several times a year)
- ☐ Occasionally (several times a month)
- ☐ Frequently (several times a week)
- ☐ Very frequently (several times a day)

How often do you look at complex Word Clouds, like the one below:

- ☐ Never
- ☐ Rarely (several times a year)
- ☐ Occasionally (several times a month)
- ☐ Frequently (several times a week)
- ☐ Very frequently (several times a day)

Appendix D: Post-Study Questionnaire

Please rate how strongly you agree or disagree with each of the following statements with respect to the tasks you completed.

1. I found bars (that depict the frequency of the words) useful when performing tasks.

- ☐ Strongly disagree
- ☐ Disagree
- ☐ Neutral
- ☐ Agree
- ☐ Strongly agree

2. I found the size of the words useful when performing tasks.

- ☐ Strongly disagree
- ☐ Disagree
- ☐ Neutral
- ☐ Agree
- ☐ Strongly agree

3. I found the color saturation of words useful when performing tasks.

- ☐ Strongly disagree
- ☐ Disagree
- ☐ Neutral
- ☐ Agree
- ☐ Strongly agree

4. Rate the three types of encoding from the easiest (starting from the left) to the most difficult (on the right):

- ☐ Bar, Color, Size
- ☐ Bar, Size, Color
- ☐ Size, Color, Bar
- ☐ Size, Bar, Color
- ☐ Color, Size, Bar
- ☐ Color, Bar, Size