CROWDSOURCING GRAPHICAL PERCEPTION OF TIME-SERIES VISUALIZATION ON MOBILE PHONES

by

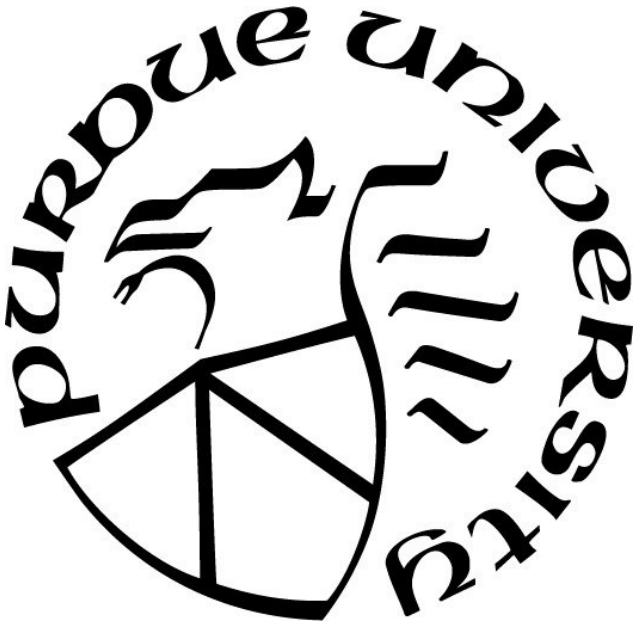
Myeonghan Ryu

A Thesis

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the Degree of

Master of Science



Department of Computer Graphics Technology

West Lafayette, Indiana

December 2020

THE PURDUE UNIVERSITY GRADUATE SCHOOL

STATEMENT OF COMMITTEE APPROVAL

Dr. Paul C. Parsons, Chair

Department of Computer Graphics Technology

Dr. Austin L. Toombs

Department of Computer Graphics Technology

Dr. Brandon J. Pitts

Department of Industrial Engineering

Approved by:

Dr. Nicoletta Adamo-Villani

Head of the Graduate Program

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

As the ubiquitous computing with mobile smart devices equipped with powerful hardware and software has become prevalent within the last decade, the demand for visual access to data has already become part of the life of many people, which requires employing the appropriate visual representation of data. Reports show that more than half of the internet traffic is through mobile, so that it is now more than only a supplemental way of desktop computers. Among many different types of data, time-series data is one of the most common types of data, such as in many news websites, personal health tracking applications, weather forecasting applications, and finance applications. Though there already exists a large body of literature on information visualization, the unique properties of mobile devices, such as the small size of the display and various context of use, make simply applying existing visualization techniques that were meant for large displays to mobile phone displays difficult. These are challenges against fully leveraging visual access to data using mobile phones. In this study, the performance with visualization on mobile phones is investigated. For this purpose, this study compares the performance of users using the two different visualization techniques to represent a collection of time-series data in limited space: line charts and horizon graphs. Methodologically, this study employs the crowdsourcing technique using Prolific (https://www.prolific.co).

CHAPTER 1. INTRODUCTION

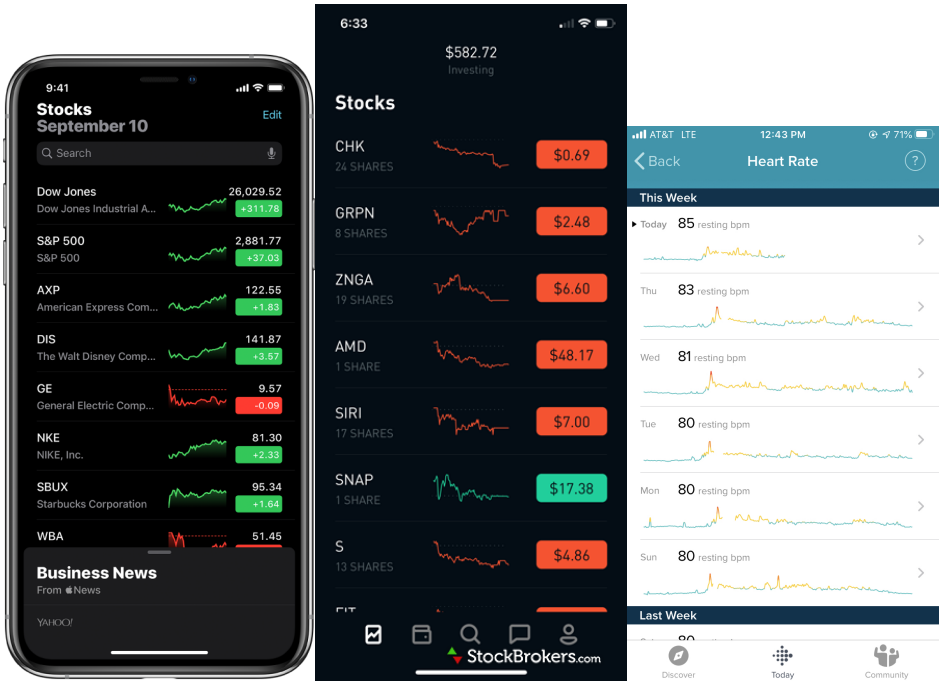
This chapter overviews the background of the problem addressed in this study and the significance of this study. Based on the overview, the research questions and hypotheses, and the scope of this study is suggested.

1.1 Background of the Problem

As the ubiquitous computing with mobile devices equipped with powerful hardware and software has become prevalent for the last decade, the demand for visual access to data through mobile devices has already become an essential part of the life of many people. The mobile platform is now one of the primary ways of accessing data and is more than only a supplemental way to access data. By Clement (2019), 52.2% of global website traffic was made through mobile phones. Also, we are witnessing the proliferation of many everyday mobile applications, such as finance (Apple, n.d.-c; Robinhood Markets, n.d.), personal health tracking (Apple, n.d.-b; Fitbit, 2015), weather (Apple, n.d.-d; Limited, n.d.)), or navigations (Google LLC, n.d.). Notably, the use of personal health tracking applications shows the critical aspect of using mobile devices: mobile devices as a primary platform for generating and consuming data. This trend requires employing the appropriate visual representation of data. Currently, mobile data visualizations are used in many everyday applications in our daily life, such as the examples above and many data journalism websites, including *The Upshot* (New York Times, n.d.) and "FiveThirtyEight" (*FiveThirtyEight*, n.d.). As this trend shows, the demand for visual access to data via mobile devices is rapidly growing. Furthermore, the necessity of adequately designed mobile data visualization is even more emphasized when visual access to the data becomes directly relevant to health and safety against disasters, such as the recent coronavirus pandemic. In a survey of lower-income parents on issues related to digital connectivity, Rideout and Katz (2016) suggested that 23 % of the families with 6- to 13-year-olds and with lower income can

access the internet only using mobile devices, and this statistic goes even higher to 34% for Hispanic and to 41% for Hispanic immigrants within the group. For these people, mobile data visualization can be the only option to visually access the data about the pandemic.

One of the most common types of data is time-series, which are sets of values changing over time (Heer et al., 2009). Because time is one of the most foundational dimensions in human life, understanding time-series data is at the root of understanding phenomena in many areas, such as finance, science, and public policy (Heer et al., 2009). Therefore, visualizing time-series data has been one of the central problems in the information visualization field. There are examples of mobile time-series visualization. Many mobile data visualizations in personal tracking apps show time series data related to a user's health, such as sleep time, activity, or vitals (Apple, n.d.-b). Weather applications (Apple, n.d.-d; Limited, n.d.) shows the changes of the data related to weather over time, such as temperature, humidity, or air quality. Specifically, visualization feature is one of the critical features in financial applications, such as stocks trading (Apple, n.d.-c; Robinhood Markets, n.d.) (See Figure 1.1), because users of these applications make decisions based on the trend of stock prices based on the visualization. Considering the prevalence of time series data, its visualizations, with its many use cases in mobile phones, time-series visualization is a crucial part of mobile data visualization.



(a) Apple Stocks app

(b) Robinhood: Invest. Save. Earn. app

(c) Fitbit

Figure 1.1. Examples of Multiple Time-series in Financial Apps on iOS (adapted from (a) (Apple, n.d.-c), (b) (Robinhood Markets, n.d.), and (c) (Fitbit, 2015))

1.2 Significance of the Study

Despite the trends described above, mobile data visualizations as a research topic have not gained much attention so far and only recently started to draw strong interests from researchers (Brehmer, Lee, Isenberg, & Choe, 2019b; Lee et al., 2018; Watson & Setlur, 2015). Though there already exists a considerable body of literature in the information visualization field, they have mainly assumed using the visualization on desktop displays. Because mobile phones have crucial differences in conditions to design and use information visualization from those for desktop display, additional research efforts are required to fully appropriate the potential of mobile phones as an environment to use information visualizations. Though the techniques for representing the relatively large amount of data on small resolution of display has been one of the

main problems of information visualization field since its beginning (Bederson, Clamage, Czerwinski, & Robertson, 2004; Sarkar & Brown, 1992; Walker, Borgo, & Jones, 2016), applying and scaling the suggested solution to mobile devices need careful consideration because of the unique properties of mobile displays (Chittaro, 2006b).

The problem the users face while using information visualizations on mobile phones emerges because mobile phones have a smaller screen with different ratios and are used in various contexts. Though effectively using a limited amount of space has been one of the primary problems in the information visualization field (Munzner, 2014), this aspect is even more emphasized for mobile data visualization. "Fat finger problem" (Siek, Rogers, & Connelly, 2005), which is caused by the size of fingers relatively bigger to the size of the display, by the main interface of visualization using thumbs causes the problem that desktop visualization users who use the small mouse pointer to interact with a visualization do not experience. Also, in most cases, the researchers of information visualization for desktop settings can expect their users to use the visualization sitting in front of the display of the computer while concentrating on their tasks because experiments were typically conducted in the lab settings. This cannot be, however, expected to the mobile data visualization users. Therefore, the research effort considering these conditions are required to reveal and resolve the challenges from these different contexts of use (Brehmer, Lee, Isenberg, & Choe, 2019a; Brehmer et al., 2019b; Lee et al., 2018; Schwab et al., 2019).

More specifically, a large number of time-series visualization use cases are found. Reflecting on its ubiquity, there already have been various approaches to visualize time-series data, including simple line charts (line charts sharing same space), line charts (using split space for each line), stacked graphs, animation, and horizon graphs (HG) aiming at improving the graphical perception of visualization users when the data density is increased (Heer et al., 2009; Javed, McDonnel, & Elmqvist, 2010; Perin, Vernier, & Fekete, 2013). Therefore, it is beneficial to examine whether the time-series visualization techniques are applicable to the mobile data visualization and how users experience the visualizations using those techniques. Among those techniques, line charts using split space for each line (LC) (Heer et al., 2009; Javed et al., 2010; Perin et al., 2013), and HG (Few, 2008; Heer et al., 2009; Reijner et al., 2008; Saito et al., 2005) are an interesting pair of visualization techniques in that they both split the space for each

time series to lower the visual clutter (Javed et al., 2010) as graphs applying small multiples technique (Tufte, 2001). However, LC more preserves the original time series, whereas HG manipulates the shape of them by coloring, layering, mirroring, and wrapping bands (Javed et al., 2010) to save even more space or draw more time-series than LC. This contrast can make differences in the performance of visualization viewers when multiple time series are represented on a mobile display.

Considering the limitation of the size of a mobile phone display, it is certainly a possible use case that scrolling on display is required to represent many time series using LC. When the number of time series is large enough, using HG also requires scrolling on display to use more space than a single display. However, by mirroring and wrapping bands, HG can increase data density more than LC without requiring interactions such as scrolling.

1.3 Research Questions

The purpose of this study is to examine whether HG can be an alternative of LC for visualizing time series data on mobile phones. To achieve this purpose, the following research questions are answered:

1. What are the performance differences between HG and LC on mobile phone display?
    (a) What are the differences in task accuracy between HG and LC on mobile phone display?
    (b) What are the differences in task completion time between HG and LC on mobile phone display?

1.4 Hypotheses

The following hypotheses are tested. In hypotheses, the small, medium, and large datasets consist of six, twelve, and twenty-four time series for each. The size of the small dataset was decided as the size not requiring vertical scrolling for LC, based on the real use cases (See Figure 1.1). Considering that the test program does not need additional elements such as search bar and

date (Figure 1.1 (a)), total investment (Figure 1.1 (b)), or button for navigating different pages (Figure 1.1 (c)), six time series for a screen looks reasonable choice for this study. Twelve time series for medium size dataset is the number of time series HG can be displayed using the same amount of space with LC. To visualize twelve time series, LC uses two pages of screen with vertical scrolling, whereas HG uses only one page of screen without vertical scrolling. Twenty-four time series for large size dataset is the number of time series HG can be displayed in two pages with vertical scrolling. To visualize twenty-four time series, LC uses four pages of screen using vertical scrolling.

H1, H2 With the small dataset, HG's advantage in showing small variations will not lead to significant performance differences, as HG requires time to mentally unstack. Therefore H1 and H2 are suggested.

H1. Using the small dataset, without any scrolling, participants will perform tasks faster with LC.

H2. Using the small dataset, without any scrolling, there will be no significant difference in task accuracy between LC and HG.

H3. Using the medium dataset without any scrolling, HG will outperform LC, with higher accuracy and less completion time. Though mentally unstacking HG takes time, it is expected that it will take more effort to read the LC when they cannot show small variations of the data because of the limited height.

H4. Using the medium dataset, when vertical scrolling is needed for LC but not HG, HG will be more effective than LC. That is, participants will do the given tasks faster with fewer errors when they use HG. When vertical scrolling is used in LC, participants should consult their working memory, whereas HG can visualize the same amount of data without using vertical scrolling so that the participants can directly compare given graphs. Though reading HG requires mentally unstacking and using eyes to switch, it is expected to have a lower cognitive load than consulting memory.

H5. Using the large dataset, where vertical scrolling is used in both cases, HG will be more effective than LC. That is, participants will do the given tasks faster with fewer errors when

they use HG. Though reading HG requires mentally unstacking, using eyes to switch, and consulting memory for two pages with vertical scrolling, it is expected to have a lower cognitive load than that of LC consulting memory for four pages of line charts.

For H3, H4, and H5, though there is no previous study directly comparing the cognitive load of mentally unstacking HG and consulting memory to compare multiple LC or small size LC, the limitation of visual working memory is already well known and emphasized in the visualization community (Munzner, 2014; Ware, 2019). Taking the popularity of HG as a successful graph design for multiple time series (Heer et al., 2009; Jabbari, Blanch, & Dupuy-Chessa, 2018a, 2018b; Javed et al., 2010; Perin et al., 2013), it was expected that the cognitive load of using HG would be less than that of consulting memory.

1.5 Delimiters

This research covers only the visualizations on mobile phones, and other kinds of mobile devices such as tablets and smartwatches are out of the scope of this study. As the resolution and computing power of tablets has been increased, tablets support the use of keyboard and mouse, and many laptops now support touch interface, the distinction between the use of tablets and laptops has become less clear than before. Also, as tablets better support pen-based interaction than mobile phones with a large display, which lessens the "fat finger problem" (Siek et al., 2005), the visualization on tablets should be independently studied. Furthermore, the way to interact with mobile devices is limited to the participants' finger, not allowing to use a touch pen.

Also, smartwatches have a very small display, which severely limits the space for representing and interacting with the data. Also, considering the field study that most of the interactions with smartwatch are quickly glancing or peeking, which often take less than 5s (Blascheck, Besançon, Bezerianos, Lee, & Isenberg, 2019; Pizza, Brown, McMillan, & Lampinen, 2016), mobile visualization use on mobile phones and smartwatches are so different that they should be independently studied.

1.6 Assumptions

For this study, it is assumed that the cultural and geographical background of the participants do not affect the performance of using visualizations on mobile devices when they have access to mobile phones.

Also, the crowdsourcing experiment participants recruited using Prolific (https://www.prolific.co), a crowdsourcing platform for conducting online research, is expected to be distracted while they are participating in the experiment because they use their mobile phones in the environment that cannot be controlled by the researcher. Also, it is assumed that among the recruited participants, some might try to get paid without giving truthful answers. Besides, the chances of participants' trying to interact with the test program against the guideline or the rule of the experiment are taken into account. Considering these expected behaviors of the experiment design, the measures to filter out the malicious participants or control or make participants stay focused are used.

1.7 Limitations

There is a limitation of using conducting an online experiment using Proflific instead of a controlled lab experiment. Though Prolific has better subject pool than recruiting participants in university (Palan & Schitter, 2018; Peer, Brandimarte, Samat, & Acquisti, 2017), a typical way of recruiting participants for experiments in literature, it's hard to control the context or experiment environment of participants, including the brightness of devices or environment, noise level affecting the participants' concentration on the test, and other distractions. This might cause the lowering of the internal validity of the study because it is hard to discern which variable affects the performance. However, these contexts that are hard to control can be advantageous in that it increases the ecological validity of the results because the diverse conditions and distractions reflect on the real use context of mobile devices.

1.8 Definitions of Key Terms

Information visualization: “The use of computer-supported, interactive, visual representations of data to amplify cognition” (Card, Mackinlay, & Shneiderman, 1999)

Time-series data: sets of values changing over time (Heer et al., 2009)

Crowdsourcing: “the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call.” (Howe, 2006)

CHAPTER 2. REVIEW OF LITERATURE

In this chapter, the relevant literature will be examined in the following order of sections: (1) Visualization for mobile phones, (2) Visualizing multiple time series data, (3) Visualization evaluation study. In the first section, the limitations and design opportunities for visualization on mobile phones will be described. Though there are many papers on mobile data visualization for PDAs, tablets, or smartwatches, only the studies about information visualization on mobile phones will be reviewed because other types of mobile devices are used under quite different use contexts and have different conditions from mobile phones for displaying information, including different screen ratios. In the second section, the techniques used for visually representing the multiple time series data will be covered. This section starts from the systematic view of visualizing time-series visualization, followed by visualizing multiple time series using small multiples and horizon graphs. In the third section, the body of literature on visualization evaluation study will be covered. Research on visualization evaluation has a considerable body of literature since the beginning of this field and is a typical approach to the visualizations. Also, as a relatively new research method for visualization evaluation study, using crowdsourcing for visualization research will be examined. Though using crowdsourcing, such as Amazon's Mechanical Turk, has various advantages in conducting visualization research, there still are some points to be noted for applying the techniques to the research.

2.1 Visualization for Mobile Phones

Even before Apple's iPhone first came out in 2007, Chittaro (2006b) paid attention to the value of mobile devices as a platform for information visualization. Discussing the disadvantages and limitations of visually representing data on mobile devices, Chittaro (2006b) indicated those restrictions: the display with a smaller size with lower resolution and fewer colors, the aspect ratio different from typical 4:3, less powerful computing power and slow network connectivity, the input peripherals inappropriate for complicated tasks, and various form-factor, performance,

and different input equipment among different mobile devices. For these reasons, he concluded that visualization for desktop computers do not easily scale to mobile devices. The typical solutions for the presentation problem, such as providing overview and detail simultaneously or representing focus with context altogether, cannot easily be applied for mobile visualization (Chittaro, 2006b). Moreover, the extremely various context of use distracts users from keeping paying attention to the visualization on the device (Chittaro, 2006b).

However, the technological development of modern mobile phones and the technological environment lessened some challenges above while leaving others still relevant. Modern mobile phones in current years are equipped with a high-resolution display, powerful computing power, and fast network connectivity, though the restrictions by the small size of display and "fat finger problem"(Siek et al., 2005) still exist. This development of mobile devices has been followed by the high demand for visually accessing the data on mobile phones. Reflecting on this trend, the body of literature examining the research space of mobile visualization in current status (Watson & Setlur, 2015) and trying to set a research agenda (Lee et al., 2018) have emerged (Brehmer et al., 2019b). The works in the literature include the evaluation of visual encoding and interaction techniques through touch interface (Brehmer et al., 2019a, 2019b; Lee et al., 2018; Watson & Setlur, 2015). More specifically, diverse data types such as time series (Brehmer et al., 2019a, 2019b), ranges (Brehmer et al., 2019b), hierarchical graphs (Brehmer et al., 2019a; Horak & Dachselt, 2018) on mobile phones were studied (Brehmer et al., 2019a). Also, some researchers approached in terms of diverse applications of the visualization on mobile phones, including health data (Brehmer et al., 2019a; Chittaro, 2006a; Dalton, Katz, & Keynes, 2018; Nicolalde, 2018; Ongwere, Connelly, & Stolterman, 2018), public transportation (Kay, Kola, Hullman, & Munson, 2016), and collaborative work using mobile data visualization over multiple devices (Badam, 2018).

2.2 Visualizing Multiple Time Series Data with Small Multiples and Horizon Graphs

In this section, the literature on visualization techniques for multiple time series data is covered. Before directly going into visualizing multiple time series, a categorization schema for visual methods to analyze time-oriented data is introduced as a means of a systematic approach to

the topic. Using this systematic categorization, the position of this study in time-series data visualization research is defined. Also, the pair of visualization for comparison in this study, small multiples and horizon graphs, can be understood in terms of this categorization. More specific studies follow this systematic categorization in the literature. The challenges in designing time series visualization and the techniques to overcome such challenges are also presented. By reviewing the literature, the basis of choosing a set of visualization techniques compared to this study will also be suggested.

2.2.1 Visualizing Multiple Time Series Data

2.2.1.1 Visualizing Time Series Data

Time series data is one of the most common types of data people face in daily life. Finding the evolution of the data over time and temporally recurring patterns has been one of the common analysis tasks in industry, academia, or even when understanding personal data, where time-series visualization can have an important role to help understand the data about themselves. The use of time series visualization even goes back to the 18th century. The seminal work of Playfair (Tufte, 2001) is a well-known example of time series visualization. Because of the prevalence of time series data and its use, there is a large body of literature on time series visualization.

Though there have been many studies on visualizing time series data, most of them were focused on specific analysis problems (Aigner et al., 2007). So, Aigner et al. (2007) developed a systematic view of the visualization of the time series data and suggested a framework to categorize visualization techniques for time-oriented data. The categorization schema they developed in their research is the following Table 2.1. They suggested three categorization criteria: time, data, and representation. And temporal primitives, and structure of time are suggested as the subcategories of time. Temporal primitives are about the composition of the time axis (Aigner et al., 2007). Timepoint is an instance in time, whereas time interval has an extent specified by two timepoints or a time point with a duration (Aigner et al., 2007). Another subcategory of time is the structure of time: linear, cycle, or branching (Aigner et al., 2007). Linear time "corresponds to our natural perception of time as being a (totally or partially) ordered

collection of temporal primitives" (Aigner et al., 2007), whereas cyclic time is "composed of a finite set of recurring temporal primitives (e.g., the seasons of the year)." (Aigner et al., 2007). A pair of examples are found in the study of (Brehmer et al., 2019b), which compared the visual representations of temperature range charts and sleep time data using bars in linear layout and radial layout. Lastly, the branching time-axis is "a split of the time axis into alternative scenarios, which is particularly relevant for planning or prediction." (Aigner et al., 2007).

As the subcategories of data, frame of reference, number of variables, and level of abstraction are suggested (Aigner et al., 2007). The frame of reference refers to the context of the data: the data collected in a non-spatial context vs. spatial data. The data in a non-spatial context means the collected data itself does not have an inherent spatial layout, whereas spatial data does. (Aigner et al., 2007). The number of time-dependent variables includes univariate and multivariate (Aigner et al., 2007). For example, a line graph with time as the x-axis and specific values as the y-axis are visualizing univariate data because it only shows the temporal changes of y values. However, a famous scatter plot used in Gapminder Trendalyzer by Rosling (2006) shows the temporal changes of two variables on the x-axis and y-axis each. Unlike a line graph with time on the x-axis, time is represented using animation by the movement of bubbles on x, y coordinates. In this case, the correlation between the data on the x-axis and the y-axis over time can be presented. The comparison of these cases shows the difference between univariate and multivariate visualizations. Level of abstraction is a subcategory of data and used when the large dataset makes it hard to represent all the data at once causing the overcrowded and cluttered displays (Aigner et al., 2007). To overcome such problems, the aggregation of data (Aigner et al., 2007; López, Snodgrass, & Moon, 2005), overview + detail interfaces (Aigner et al., 2007; Schneiderman, 1996), and feature visualization (Aigner et al., 2007; Silver, 1994) have been suggested.

Lastly, the representation contains time dependency and dimensionality as its subcategories (Aigner et al., 2007). Time dependency distinguishes static and dynamic representation of time-oriented data, which is whether the representations change over time (Aigner et al., 2007). For example, a typical line graph is a static representation, whereas the Gapminder Trendalyzer (Rosling, 2006) using an animation technique to show the temporal change of the values on the x-axis and y-axis is a dynamic representation of data. Moreover, the

dimensionality–2D or 3D–is another subcategory of the representation (Aigner et al., 2007). It is widely accepted that 3D should be used with careful considerations and enough justification (Munzner, 2014). Except for the cases with strong justification, it is more desirable to use 2D than 3D (Munzner, 2014). However, because there are some cases certainly requiring the use of 3D, like flow or volume data, and there are advanced interaction techniques or additional visual cues against the disadvantages of using 3D, it should be considered as an option depending on task and data (Aigner et al., 2007).

This categorization scheme will be used for figuring out a pair of visualization techniques evaluated in this study, small multiples, and horizon graphs, after examining them in the following sections.

Table 2.1. *Categorization Schema for Visual Methods for Analyzing Time-oriented Data (Aigner et al., 2007)*

Time	Temporal primitives	time points	time intervals
	Structure of time	linear cyclic	branching
Data	Frame of reference	abstract	spatial
	Number of variables	univariate	multivariate
	Levels of abstraction	data	data abstractions
Representation	Time dependency	static	dynamic
	Dimensionality	2D	3D

2.2.1.2 Visualizing Multiple Time Series Data

When visualizing time series data, the problem of clustering, occlusion, or over-plotting happens with the large dataset containing lots of data items. For instance, when visualizing the daily temperature data over ten years using a line graph, the data would contain about 3,650 data items, 365 for a year, which is too many to represent using x, y coordinates without clustering or occlusion. Walker et al. (2016) categorized and evaluated the techniques to resolve such an issue. They categorized the techniques as data aggregation, lens-based approach, and layout based approach (Walker et al., 2016). The data aggregation is aggregating data points in a meaningful way, depicting "statistical features of the items in each segment of time through a meaningful

visual mapping" (Walker et al., 2016). Controlling the granularity of the time, such as day, week, or month, fall into this type of technique. The lens-based approach is adding in-place magnification to distort the time axis to represent specific parts more in detail while maintaining the context (Walker et al., 2016). Usage of this technique can be found in the daily use of mobile devices, like the zooming feature in many applications. Layout based approach is modifying "the spatial arrangement of the time-series to provide a linear mapping of time while transforming time-series graphs" (Walker et al., 2016). According to Walker et al. (2016), this type of technique resolves the over-plotting issue by stacking graphs with different levels of zooming on demands (Javed & Elmqvist, 2010) or by simultaneously providing overview and detail displays (Plaisant, Carr, & Shneiderman, 1995).

However, these techniques cover only the cases where the time axis is too much crowded. When drawing multiple line graphs on the same x, y coordinates, the over-plotting problem happens with a much smaller number of data items. Javed et al. (2010) addressed this issue. Before Javed et al. (2010), most studies about the performance of multiple line graphs for time series data involved only two time-series, and it was not certain whether the results from the studies comparing two line graphs could be generalized (Javed et al., 2010). However, considering that the tasks involving many time series simultaneously is common (Hochheiser & Shneiderman, 2004; Javed et al., 2010), they evaluated and suggested the results of the experiments covering the techniques to visualize multiple time series (Javed et al., 2010). The techniques include small multiples (Tufte, 2001), stacked graphs (Byron & Wattenberg, 2008), horizon graphs(Saito et al., 2005), and braided graph (Javed et al., 2010). Based on the results of experiments with the tasks to "find the time series with the highest value at a specific point in time" (Javed et al., 2010; Lam, Munzner, & Kincaid, 2007), "to find the time series with the highest increase during the whole displayed time period" (Beattie & Jones, 2002; Javed et al., 2010), and "to find the individual values of each time series and then figure out which one was the largest one" (Javed et al., 2010; Simkin & Hastie, 1987), they found that "shared space techniques (simple line graph and braided graph) were faster than split-space techniques for the local Maximum task, split-space techniques (small multiples and horizon graphs) were faster than shared-space techniques for the dispersed Discrimination task, and the slope task, with a dispersed visual span, was special–small multiples and simple graphs were faster here "(Javed et al., 2010).

2.2.2 Visualizing Time Series Data Using Small Multiples

Small multiple is a visualization technique making "the same design structure repeated for all the images" (Tufte, 1990). It is an economical way to represent multivariate time series data because when visualization viewers understand a slice, then they can directly access to all slices by an only slight movement of their eyes (Tufte, 1990). Because of constancy in design structure, "the design allows the viewer to focus on changes in the data rather than on changes in graphical design" (Tufte, 1990). Therefore, small multiples are inherently comparative and multivariate (Tufte, 2001).

Because of its effectiveness for representing multivariate data and for comparative tasks, it has been widely used as a visualization technique supporting comparison tasks and dealing with clutter or occlusion problem (Robertson, Fernandez, Fisher, Lee, & Stasko, 2008; Tufte, 1990). Currently, these small multiples designs are one of the prevalent visualization techniques on mobile phone applications with visualization features (Brehmer et al., 2019a), such as personal activity or health applications(Apple, n.d.-a, n.d.-b; Fitbit, 2015) or finance apps (Apple, n.d.-c; Robinhood Markets, n.d.), just name a few. Also, considering the description about small multiple by Tufte (1990) with many cases in the literature, this technique can be used with diverse representations such as typical line graphs (Javed et al., 2010), horizon graphs (Few, 2008; Reijner et al., 2008), scatter plot (Brehmer et al., 2019a; Robertson et al., 2008), animation with traces(Brehmer et al., 2019a; Robertson et al., 2008), data glyphs (Fuchs, Fischer, Mansmann, Bertini, & Isenberg, 2013), graph data structure or network (Archambault, Purchase, & Pinaud, 2011), flow maps (Boyandin, Bertini, & Lalanne, 2012), or even a dashboard with multiple types of visual representations (Ondov, Jardine, Elmqvist, & Franconeri, 2019).

Though small multiples designs can be used to represent multivariate data or multiple time series without the clutter or occlusion problem, it was indicated that the size of each slice decrease as the number of slices increases, making the visualization less readable (Robertson et al., 2008). Considering this aspect with mobile devices with more limited screen resolution than desktop, it can be more critical. Robertson et al. (2008) indicated that this limitation does not mean that the small multiples technique represented a more limited number of data points than animation or trace. Because when the size of each display becomes very smaller as the number of

data points increases, the clutter also becomes extreme in animation or trace (Robertson et al., 2008). However, it is unclear whether this point is generalizable to mobile phone display because it is much smaller than the desktop monitors, and it has a different screen ratio.

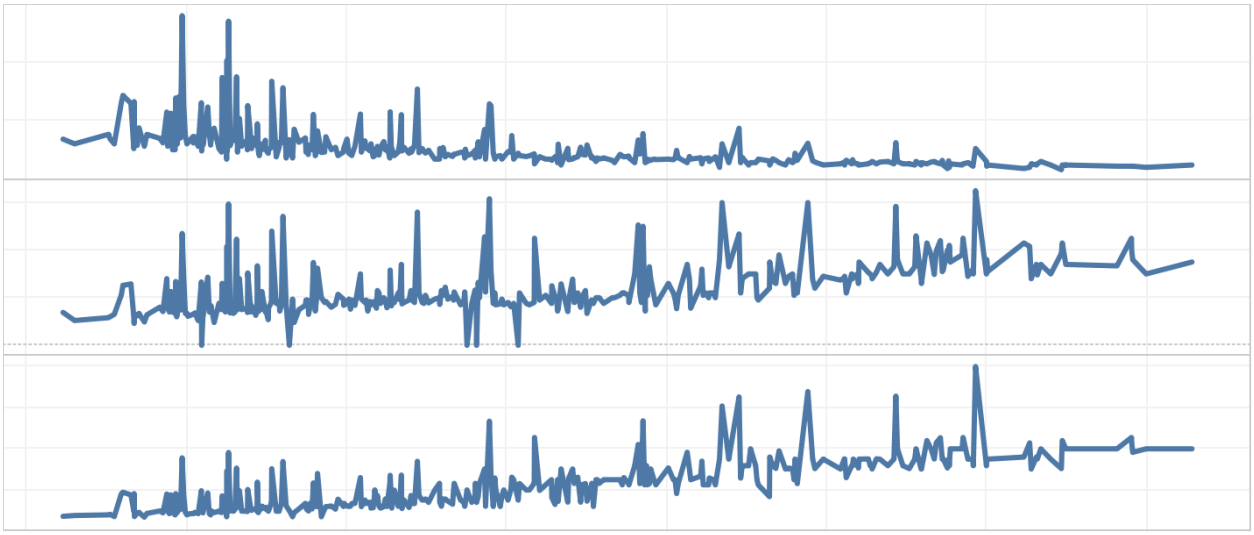


Figure 2.1. Small Multiple Line Graphs

2.2.3 Visualizing Time Series Data Using Horizon Graphs

Horizon graph was first introduced as "Two-Tone Pseudo Coloring" by Saito et al. (2005) and has been recognized as a modern visualization technique to visualize multiple time series in the research community (Federico, Hoffmann, Rind, Aigner, & Miksch, 2014; Gogolou, Tsandilas, Palpanas, & Bezerianos, 2019; Heer et al., 2009; Jabbari et al., 2018a; Javed et al., 2010; Perin et al., 2013). In industry, Panopticon (Few, 2008; Reijner et al., 2008) commercialized this technique and gave an example of visualizing a large number of data (e.g., daily changes in the 50 stock prices) in a spatially efficient way with reasonable cognitive costs while properly supporting to detect anomalies, to represent each item independently, to compare between items, and to visualize changes requiring further examination precisely. It can be seen as well reflecting on the famous mantra in this field by Ben Shneiderman, "Overview first, zoom and filter, then details-on-demand" Schneiderman (1996). The cost of mirroring negative values, dividing multiple time series into independent bands, and layering for retaining more vertical

space for more time-series requires mental unstacking, mentally re-drawing the graph in the simplest form. But Few (2008) insisted that the advantage of freeing more vertical spaces while effectively supporting such tasks worth the cost of the mental unstacking.

The research effort to scientifically examine the advantages and disadvantages of using horizon graphs to visualize multiple time series has been for a decade. Heer et al. (2009) examined the points with a set of controlled experiments for value comparison tasks. They conducted experiments to evaluate the effect of the number of bands, mirroring and layering in a horizon chart, and the chart size to estimation time and accuracy. They suggested that "mirroring does not hamper graphical perception" (Heer et al., 2009), and "layered bands are beneficial as chart size decreases" (Heer et al., 2009). However, they only examined the value comparison task of a horizon graph. Javed et al. (2010) conducted experiments comparing the techniques to visualize multiple time series, including simple line graph, braided graph, small multiples, and horizon graphs, with more tasks: "to find the time series with the highest value at a specific point in time," (Javed et al., 2010) "to find the time series with the highest increase during the whole displayed time period," (Javed et al., 2010), and "to find the individual values of each time series and then figure out which one was the largest one" (Javed et al., 2010). Perin et al. (2013) introduced Interactive Horizon Graph, the improved version of the horizon graph, adding baseline panning and value zooming interaction. Federico et al. (2014) suggested the Qualizon graph integrating qualitative information with quantitative detail over a horizon graph. Jabbari et al. (2018a) conducted a series of experiments on the performance comparison among line graphs, horizon graphs, and composite visual mappings, including position-value, hue-value, texture-saturation, and hue-saturation. Most recently, Gogolou et al. (2019) examined the similarity perception in time series visualization using a line graph, horizon graph, and colorfield by analyzing the Electroencephalography (EEG) signal.

As presented above, the horizon graph is getting much research focus over a decade because of its effectiveness to represent multiple time series data with high data density while effectively supporting detecting anomalies and comparison over many time series. Considering the ubiquity of time series data, increasing demands of visual access to those data through mobile phones, and far more limited estate of mobile phone display than desktop monitors, it is timely to explore the design opportunity of the horizon graph on a mobile display. Especially, considering

the limitation of previous works, which only covered the value comparison tasks of a single horizon graph (Heer et al., 2009), and which considered different tasks not covered by (Heer et al., 2009) but without the detail aspects of horizon graphs such as the number and the size of bands covered by Heer et al. (2009), examining such points on mobile phone display will be a meaningful contribution.

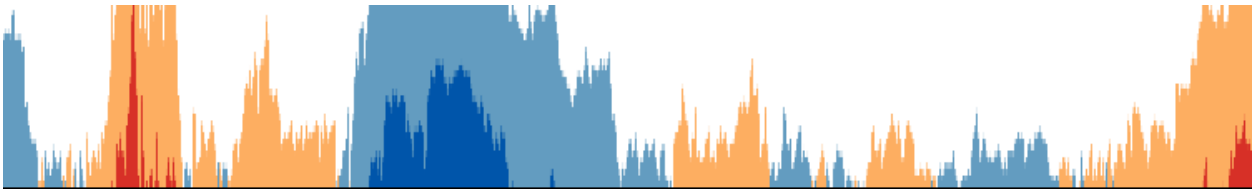


Figure 2.2. Horizon Graph

2.3 Visualization Evaluation Using Crowdsourcing

2.3.1 Visualization Evaluation Studies

This study follows the traditional approach established in the information visualization field by Cleveland and McGill (1984). In their seminal work on the graphical perception of visual elements by an experimental approach, they identified a set of "elementary tasks and perceptual tasks that are carried out when people extract quantitative information from graphs" (Cleveland & McGill, 1984), including a position on a common scale, position on non-aligned scales, length, direction, angle, area, volume, curvature, and shading. Also, they examined how quantitative information from graphs was used when people view the various charts, such as bar charts, pie charts, or scatter plots (Cleveland & McGill, 1984). And based on the results of the experiments, they ranked the tasks by the accuracy. (Cleveland & McGill, 1984). After that, these experimental methods have been used by many researchers to examine the effect of visual encoding on the accuracy or response time Heer and Bostock (2010). Most of the articles examined in this study applied this experimental approach established since Cleveland and McGill (1984).

Huang, Eades, and Hong (2008, 2009) suggested an approach to evaluate visualization in terms of cognitive load. They indicated that there could be some situations where task performance measure is not sensitive enough to show the differences of visualizations, such as when the same level of performances is measured requiring different amounts of cognitive load."Cognitive load refers to the amount of cognitive resources needed to perform a given task" (Huang et al., 2009), therefore it "can also be called 'memory demand"' (Huang et al., 2009; Wickens, Hollands, Banbury, & Parasuraman, 2015). As one of the ways of measuring cognitive load, they suggested subjective or self-report measures, one of which is reporting mental effort "by mapping the perceived amount of mental effort to a numerical value . . . from 1 to 9" (Huang et al., 2009; Paas, 1992) and is suggested as "reliable, non-intrusive and sensitive to small changes in memory demand" (Huang et al., 2009; Tuovinen & Paas, 2004).

2.3.2 Visualization Evaluation Using Crowdsourcing

Most graphical perception studies have conducted controlled experiments in a laboratory. However, as the crowdsourcing got the attention as a viable way to collect and take advantage of the labor of crowd (Borgo et al., 2017), visualization researchers have considered it as a way to experiment while overcoming the limitations of controlled laboratory studies with its scalability and low-cost (Borgo et al., 2017; Heer & Bostock, 2010). Crowdsourcing is "a new labor market phenomenon where simple, often monotonous labor tasks are replaced by open self-managed recruitment of large groups of people from the general public" (Borgo et al., 2017).

Crowdworkers are often micro-paid per small tasks they complete (Heer & Bostock, 2010). Amazon's Mechanical Turk is a famous example (Buhrmester, Kwang, & Gosling, 2011), and recently the alternatives are emerging with different target users, such as Prolific (https://www.prolific.co) for researchers and startups and Crowdflower (https://www.crowdflower.com) for companies working on data science (Peer et al., 2017). In recent years, using crowdsourcing instead of experiments under controlled lab setting is gaining much attention from visualization researchers. The annual numbers of studies using crowdsourcing for their experiments show an upward trend since 2009, when the first paper employing crowdsourcing for its visualization experiment was published (Borgo et al., 2017).

This trend emerged because of the clear advantages of using crowdsourcing. By employing a crowdsourcing experiment instead of a controlled experiment, researchers can collect many participants from larger participant pools with a much lower cost in a short time (Borgo, R., Micallef, L., Bach, B., McGee, F., Lee, 2018). As researchers can conduct their experiment with larger sample sets, they can expect greater statistical significance (Borgo, R., Micallef, L., Bach, B., McGee, F., Lee, 2018). Also, because crowdworkers are expected to have more varied backgrounds and demographics, the generalizability of the findings is assured (Borgo, R., Micallef, L., Bach, B., McGee, F., Lee, 2018). Though researchers could lose some control over the participants of the experiment, which could be a factor lowering the internal validity of the experiment, simultaneously, the ecological validity increases because less control reflects on more real context of use (Heer & Bostock, 2010).

Besides, by replicating the previous graphical perception study, which employed conventional control experiments, using crowdsourcing and comparing both results, Heer and Bostock (2010) suggested that employing crowdsourcing for an experiment is a viable alternative of the control experiment. With Heer and Bostock (2010), Borgo et al. (2017) also indicated that crowdsourcing could be used for comparative study, which is "to compare two visualization techniques in terms of their ability to support different tasks and workflows."(Borgo et al., 2017). Though there are concerns about the demographics of crowdworkers (Goodman, Cryder, & Cheema, 2013) and controlling the context participants perform the experiments (Brehmer et al., 2019b), considering larger and diverse subject pool (Borgo et al., 2017) and successful replications of existing lab experiment in the visualization field Borgo et al. (2017); Heer and Bostock (2010) and even other fields including economics and psychology (Amir, Rand, et al., 2012; Crump, McDonnell, & Gureckis, 2013; Horton, Rand, & Zeckhauser, 2011; Paolacci, Chandler, & Ipeirotis, 2010; Peer et al., 2017; Suri & Watts, 2011), using a crowdsourcing platform for a visualization graphical perception study is a reliable option.

2.4 Summary of the ROL

The generalized points from the review of the literature propose that this study is valid in its importance and methods. Time series visualization is one of the most prevalent types of

visualization in daily life, so that there have been many studies on designing spatially efficient time series visualization. Also, considering the high demand for visually accessing data on ubiquitous mobile phones and its vast share in total internet traffic, studying the way to efficiently representing time-series visualization on mobile phones is valid in its importance. Also, using crowdsourcing for an experiment is one of the proper ways to gain ecological validity and the generalizability of the results. Furthermore, methodologically, employing this emerging way of experiment in the information visualization and reflecting on the process, and the results will also be a meaningful contribution to the research community.

CHAPTER 3. METHODOLOGY

This study is aimed at comparing the task performance of LC and HG for visualizing time-series data on mobile phones. For this purpose, the graphical perception experiment comparing the visualization viewers' performance using LC and HG with different conditions was conducted using Prolific, a crowdsourcing platform for conducting online studies. Along with the static LC and HG layout, the vertical scrolling interaction is one of the crucial factors to be taken into account when visualizing a large dataset that cannot be represented on a display, maintaining readability. Notably, it is inevitable to involve the vertical scrolling for visualizing many multiple time series using LC because line graphs with too low height cannot sufficiently visualize the variation of values. Considering the real use cases (Apple, n.d.-c; Robinhood Markets, n.d.), when a user wants to add more than eight stock items in the case of Figure 1.1, the use of vertical scrolling is inevitable. In particular, because mobile phones have much less screen resolution than desktop monitors, mobile phone users more scroll than when desktop users using monitors (Kim, Thomas, Sankaranarayana, Gedeon, & Yoon, 2016). For these reasons, along with the cases without using the interaction to compare the basic graphical perception of two layouts, the cases involving the vertical scrolling interaction were considered.

3.1 Experiment Design

The experiment employed 2 x 3 x 2 factorial design, with the size of the dataset (small/medium/large), layout (LC / HG), and use of vertical scrolling (use or not) as independent variables. But as shown in Table 4.1, some conditions were not tested. Further detail is covered below in the treatment section. The experiments used the mixed-design of between-subjects design and within-subjects design, setting the layout as the between-subjects factor and others as within-subjects factors because the main focus is the comparison of LC and HG under different conditions.

Also, the experiments were conducted after pre-registering the experiment design to the Open Science Forum to make the results more valid and reliable (https://osf.io/e4rvy). Nosek, Ebersole, DeHaven, and Mellor (2018) pointed out that while scientific progress partly relies on generating and testing hypotheses, "the distinction between postdiction and prediction is appreciated conceptually but is not respected in practice." The pre-registration process is suggested to prevent a researcher from doing postdiction by registering a study before conducting an experiment. This study followed this process.

3.1.1 Data

For the experiments, the real-world data from Nasdaq (https://www.nasdaq.com/market-activity/quotes/historical) was used because, considering mobile devices' use context, the differences of the controlled lab experiments and the real-world situation can be more significant than those of using desktop or laptop in that people are more easily distracted. Also, it is easier to switch to different tasks while doing a task using a mobile phone. In addition to DOW 30 items, eleven more items were added to make sure that all the items have the data for the last five years and the graphs have clear distinctions in their shapes and trends. For every trial in every condition, a predetermined number of items were randomly chosen and visualized depending on its size of the dataset.

3.1.2 Participants

3.1.2.1 Pilot Study

For three rounds of the pilot study, three participants were recruited for the first pilot study through word-of-mouth. Fifteen and twenty participants were recruited for each of the second and third pilot studies through Prolific. Of the three first pilot participants, one was a nurse, and two were graduate students, one studying HCI and another one studying Social sciences. Two of them were not familiar with visualization. Their ages were the late 20s or early 30s. Those were recruited considering the report of demographics of participants on Mechanical Turk that they

Table 3.1. *Experiment Design*

		Layout	
		LC	HG
Vertical scrolling	Not used	Size of data set small / medium / ~~large~~	Size of data set small / medium / ~~large~~
	Used	Size of data set ~~small~~ / medium / large	Size of data set ~~small~~ / ~~medium~~ / large

tend to be more educated (Goodman et al., 2013), this population was appropriate for the pilot study. Though the report also indicated that the participants tend to be Caucasian/European Americans (Goodman et al., 2013), the racial differences do not look to have a considerable effect on the performance of the tasks. More details about the pilot study are covered in another section.

3.1.2.2 Primary Study

For the primary study, 196 participants were recruited through Prolific. The number was determined by running the a priori power analysis based on the second and third pilot studies' results ran on Prolific. The effect size of the response time was $Cohen's\ d = \frac{Mean_{lc\ response\ time} - Mean_{hg\ response\ time}}{\sqrt{\frac{(n_{lc}-1)*\sigma^2_{lc\ response\ time} + (n_{hg}-1)*\sigma^2_{hg\ response\ time}}{n_{lc}+n_{hg}-2}}} = \frac{7648.41-5006.29}{2057.19} \approx 1.28$. And the effect size of the accuracy was $Cohen's\ d = \frac{Mean_{lc\ accuracy} - Mean_{hg\ accuracy}}{\sqrt{\frac{(n_{lc}-1)*\sigma^2_{lc\ accuracy} + (n_{hg}-1)*\sigma^2_{hg\ accuracy}}{n_{lc}+n_{hg}-2}}} = \frac{0.621-0.574}{0.099} \approx 0.473$. Since the d of the accuracy was smaller than that of the response time, the number of participants for the primary study had to be enough to catch the effect of layout on the accuracy. By running the a priori one-tailed power analysis with the calculated effect size, $\alpha = .05$, and the power $\beta = .95$ for the two independent means, 98 participants for each layout was calculated.

The population of participants was not strictly limited to a particular group. Considering the previous studies employing crowdsourcing on visualization research (Borgo et al., 2017; Borgo, R., Micallef, L., Bach, B., McGee, F., Lee, 2018; Brehmer et al., 2019a, 2019b; Heer & Bostock, 2010) and the advantage of using crowdsourcing to collect the participants with a more diverse background than standard internet samples or American college samples (Buhrmester et al., 2011), it was expected that the population of participants would be more representative of

possible viewers of the visualizations than those recruited for a controlled laboratory experiment at university. Therefore, the participants were limited to the ones with the age between 18 and 50, who are fluent in English to read instructions, and with the Prolific's approval rate over 95%.

3.1.3 Visualization Design

2-band HG was used following Heer et al. (2009)'s design guideline to use 2-band HG based on their results, showing better performance than 3-band or 4-band HG. Also, as a baseline, $y_{b_0} = \frac{y_M - y_m}{2}$, with y_M and y_m being the maximum and minimum values in the dataset for each graph, was used to differentiate the areas with different colors, red for below the baseline and green for the above. This color mapping was decided considering the color scheme being used in stock portfolio charts, red for decrease and green for the increase of stock price. Though the standard color scheme of horizon graphs from previous studies (Heer et al., 2009; Javed et al., 2010) and example of visualizing stock prices using horizon graphs (Few, 2008; Reijner et al., 2008) were considered at the first time, finally the color scheme with red for negative and green for positive was used after the first pilot study where the accuracy of the slope task with horizon graphs was significantly low. The possibility was considered that the participants might have been confused with the meaning of the red-blue color because one of the symbolic meanings of the red was sometimes related to something dangerous or negative, the increase of the prices is considered as good in many cases in the stock market. Also, considering the possibility where the people might have been confused with the mapping of color and positive/negative, the offset mode of the horizon graph (Heer et al., 2009) was used instead of the typical mirror mode. Because Heer et al. (2009) suggested that there were no performance differences between using horizon graphs with mirror mode and the one with the offset mode, it was expected that using the offset mode would not have at least a negative impact on the performance in this study.

Though the use cases of Apple Stocks (Apple, n.d.-c) and Robinhood (Robinhood Markets, n.d.) sometimes provide dashed lines to show the opening price of a stock item in a day, considering the purpose of the graphical perception study examining the effect of using HG, each horizon graph was drawn to appropriate the features of HG as much as possible. Also, the

baselines for splitting the area with each color into two tones, $y_{b_1} = \frac{y_M - y_{b_0}}{2}$ and $y_{b_2} = \frac{y_{b_0} - y_m}{2}$ were be used. And the participants were able to vertically scroll as much as they need.

Because this study evaluates the visual perception of time series, any additional features were not added to the graphs. As Javed et al. (2010); Perin et al. (2013), no scale, tick, and numeric values were added. The baseline that can be seen in the graph of GE in Figure 1.1 (a) was not added. But as references, numbers starting from 0 were assigned to each graph.

3.1.4 Treatment

From twelve possible experiment conditions by combining independent variables, which were the size of the dataset, layout, and use of vertical scrolling, only seven conditions were tested (See Table 3.1), excluding the cases not using vertical scrolling for the large dataset, the ones using vertical scrolling interaction for a small dataset, and the one where vertical scrolling interaction was used for HG with the medium dataset. They were not relevant, considering the goal of this study.

Table 3.2. *Ratio of Resolutions of Pairs*

	Virtual resolution of display* (LC : HG)	Real height of each graph (LC : HG)	Virtual resolution of each graph (LC : HG)
Pair 1	1:1	1:1	1:2
Pair 2	1:1	1:1	1:2
Pair 3	2:1	2:1	1:1
Pair 4	4:2	2:1	1:1

* 1 represents the resolution of single display without vertical scrolling

3.1.5 Tasks

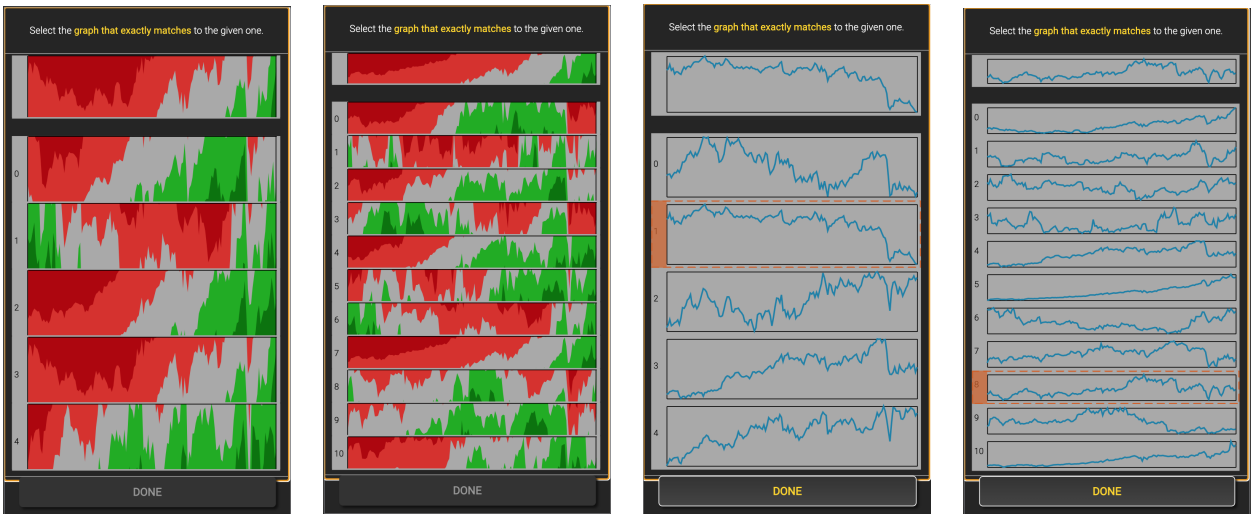
Because time-series visualization techniques have been one of the main topics in this field, there have already been efforts to evaluate the techniques for different tasks. To choose the tasks for the experiment, the taxonomy by Andrienko and Andrienko (2006) was used because the

categorization of *elementary tasks*, which deals with individual elements, and *synoptic tasks*, which consider sets of references, was found useful in analyzing the real use cases of the tasks. For example, using Apple Stocks (Apple, n.d.-c) and Robinhood (Robinhood Markets, n.d.), users can make a wish list of stock items of which the prices are represented using LC. From the list, they can choose a stock item they are interested in or want to buy or sell, and the details about the item are provided on-demand. It is a partial application of Ben Shneiderman's mantra, "Overview, zoom and filter, then details on demand" (Schneiderman, 1996). Because the use of LC in the apps provides an overview of the changes in the price of multiple stock items, most relevant tasks, in this case, are understood as the *synoptic tasks*.

Considering the taxonomy by Andrienko and Andrienko (2006), the work of Perin et al. (2013), which deployed the taxonomy for evaluating their Interactive Horizon Graph and Reduced Line Charts for the *Maximum*, *Discriminate*, and *Matching* tasks, and Javed et al. (2010) which tested the *Maximum*, *Slope*, and *Discrimination* tasks, the tasks for this study was decided.

- *Matching*

  The participants were required to choose the matching time series in its shape with given reference time series (*Synoptic task* for *relation-seeking*) (Andrienko & Andrienko, 2006). This task was chosen considering the comparison of market summary indices, such as NASDAQ or Dow Jones, with specific stock items.(Javed et al., 2010; Perin et al., 2013)



(a) HG (small dataset)

(b) HG (medium/large dataset)

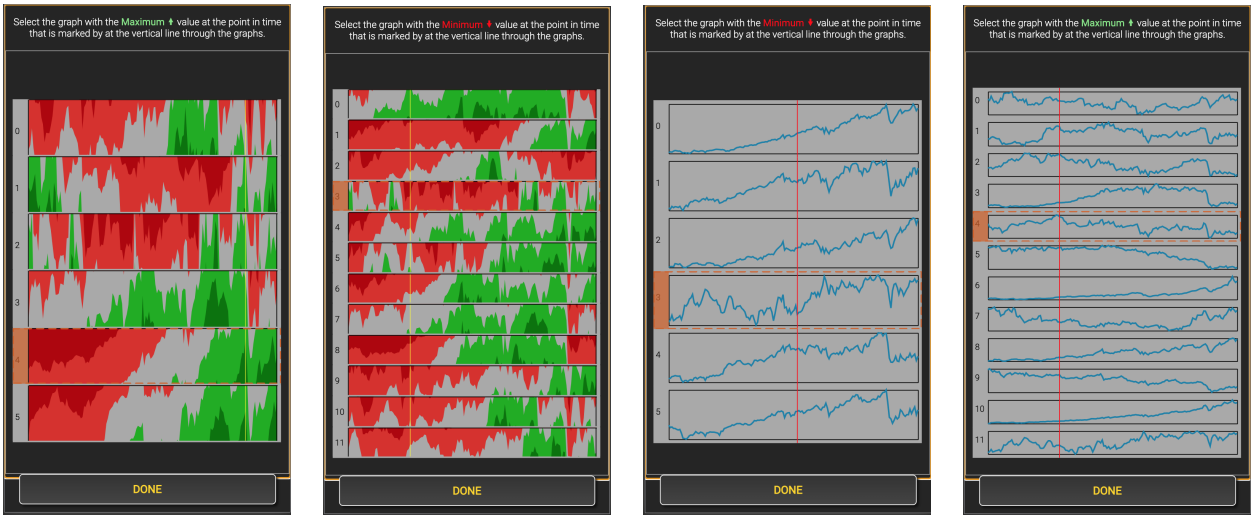
(c) LC (small dataset)

(d) LC (medium/large dataset)

Figure 3.1. Matching Task Examples

- *Maximum/Minimum*

  The participants were required to choose the time series with the highest points on its y-scale at t (*Elementary task* for *direct comparison* (Andrienko & Andrienko, 2006)). It is to test the performance of basic comparison among values in a time series (Javed et al., 2010; Perin et al., 2013).



(a) HG (small dataset)

(b) HG (medium/large dataset)

(c) LC (small dataset)

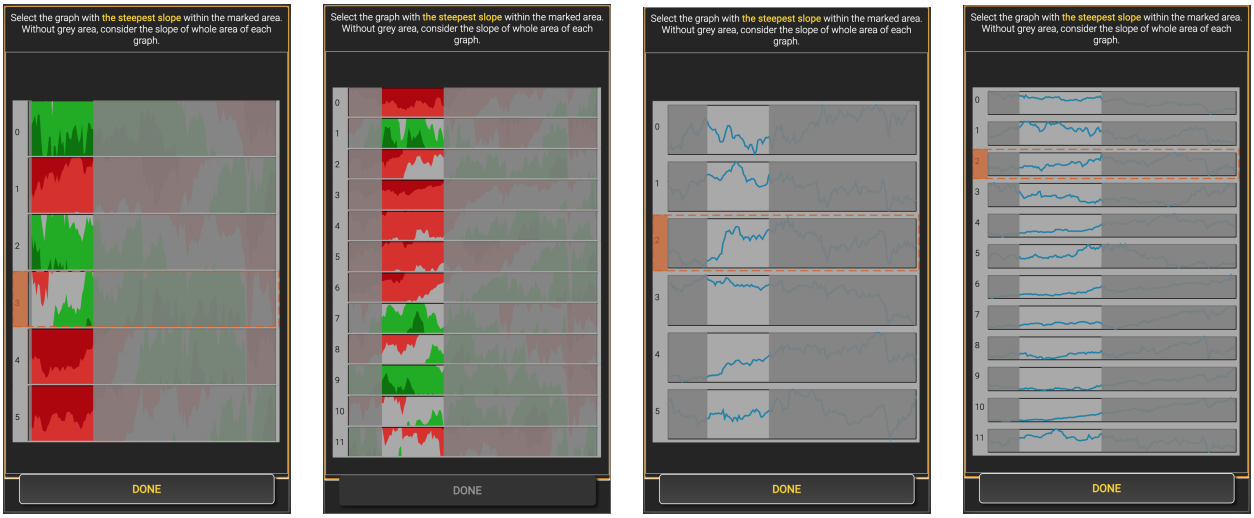
(d) LC (medium/large dataset)

Figure 3.2. Maximum/minimum Task Examples

- *Slope*

  The participants were required to compare the slopes within a given time interval. They chose a time series of which the represented price was most increased/decreased in a given time interval (*Synoptic task* for *direct pattern comparison*) (Andrienko & Andrienko, 2006). It is to make decisions on choosing which one to get more detailed information, the comparison of trends over multiple time series is essential. Javed et al. (2010) assessed *Slope* task as a global task requiring participants to find the time series with the highest increase in whole the graph, but in this study, the time interval for the slope was a part of the display, with 1/2, 1/3, 1/4 of the width of each graph. (Javed et al., 2010)

For each trial, completion time and whether the response to the trial was correct were recorded. The completion time was measured from when the participants touched the instruction about each trial to the moment when they select their response and touch the DONE button below the screen. And the accuracy was calculated by the individual based on whether each response was correct. To collect subjective responses, the participants were asked to conduct a



(a) HG (small dataset)

(b) HG (medium/large dataset)

(c) LC (small dataset)

(d) LC (medium/large dataset)

Figure 3.3. Slope Task Examples

self-assessment of their performance at the end of the trials of each condition in terms of response time and accuracy were collected and to rate these using a Likert scale from 1 (low) to 5 (high).

The experiment was designed to take approximately 25 minutes, including practice sessions. Participants received $3.96 when they finish all the tasks and submit the completion code to Prolific. When all the study was done in Prolific, the average reward rate was $10.12 per hour, which was around average in Prolific, based on the dashboard given by Prolific. The initial number of trials is two practice trials and five timed trials in LC and three practice trials and five timed trials in HG in each experiment condition, and 90 or 78 trials in total. The differences of the number of trials for LC and HG was because there were four conditions using LC and three conditions using HG. In the trials involving vertical scrolling, half of the answers were on the second page where the participant needs to use the scrolling, and the rest of the answers were on the first page, not requiring vertical scrolling. Because the chances are that when a task involves scrolling, the answer's location might affect how much attention the participants would give to each element (Guan & Cutrell, 2007; Kim et al., 2016). But the exact number of trials was changed depending on the additional practice trials each participant tried.

The order of tasks was as suggested above, because it was expected and also turned out by the pilot studies that the matching task was the easiest one, the maximum/minimum was the next, and the slope task was the hardest one. For the *Matching* task, participants could find the correct answer by skimming through the graphs to figure out overall shape. For the *Maximum/minimum* task, they needed to accurately compare the position of the highest or lowest point at the vertical line, which required more accurate reading of the graph than *Matching* task. for the *Slope* task, they needed to calculate the slope when the y values at the starting and ending points of a span were different each other, requiring much more complicated caluclation to compare with each other. Considering that the participants were expected to be general population without expertise in graph reading, they were expected to lose their focus easily if they started with the hardest one. Therefore, the experiment was designed to gradually increase the difficulty of the tasks.

As the purpose of this research does not include evaluating the visualization literacy of the participants at Prolific, practice sessions for each type of task on every visual representation were given to make participants familiar with reading the visual representations. Because HG is less common in use cases, participants might not be familiar with using it. Therefore, a brief

description of how the HG is made and how to read HG was given before the test and during the practice sessions. Two test trials for the participants using line charts and 3 test trials for the participants using horizon graphs for each condition were prepared. Also, they could practice more sessions as much as they wanted.

To ensure participants stay focused on the tasks, there was a time limit of 77 minutes to complete the tasks to get a reward, and intermittent easy tasks were inserted and required to answer correctly to get a reward (Brehmer et al., 2019b). For the matching task, the attention trial was finding a graph that matches to the example, the only one orienting downward from five graphs. For the maximum/minimum task, the participants were asked to choose the one with the maximum value at the vertical line among six graphs when only one graph had a positive value at the vertical line. For the slope task, which the results of the pilot study suggested that this task is the most difficult one, the participants were asked to select the one with the steepest slope in the span 1/4 width of the whole graph, when only one graph had a steep slope and others were relatively flat. In this case, they were directly given the index of the one with the steepest slope. This time limit and required intermittent tasks were notified before starting the tasks in Prolific. Also, to filter the fast deceivers (Borgo et al., 2017; Gadiraju, Kawase, Dietze, & Demartini, 2015), the minimum threshold for the correctness of the answers to get a reward was decided (Borgo et al., 2017) and considered with the total time taken and the average response time for each condition.

Also, considering the "fat finger problem" (Siek et al., 2005), choosing the correct value on the chart was not clicking the exact point. Instead, it was choosing an area covering the exact point of value answering multiple-choice questions, referring to the case of Brehmer et al. (2019b).

3.1.6 Implementation

The test application was a mobile web application only accessible on mobile devices with the portrait display mode. The visualization for the test application was implemented using the D3.js library (Bostock, Ogievetsky, & Heer, 2011). The server ran on the Microsoft Azure Cloud computing service (Microsoft, n.d.). The backend server for the web application was written

using Node.js (Joyent Inc., n.d.). The color scheme of the test program user interface referred to Brehmer et al. (2019a, 2019b). Also, referring to the test application of Brehmer et al. (2019b), the orientation of the display was limited to "portrait," and the size of the displayed testing program was limited using a CSS media query. The width was limited to between 320 px and 1440 px, and height was limited up to 1440 px. With this limitation, the size of the displayed testing application was limited even with the devices with a relatively large display for smartphones.

3.2 Experiment

Before the primary experiment, three pilot tests were conducted to ascertain the appropriateness of the planned procedure for expected participants, the comprehensibility of the tasks and instructions, allocating enough time, the robustness and proper difficulty of attention check, and the fairness of reward for the participants (Borgo et al., 2017). The first pilot study was conducted iteratively with three participants. They were given the link, did the tasks, and gave feedback. Reflecting on their common feedback that the experiment too long than expected, the experiment design was changed from fully within-subjects design to the mixed design of between-subjects and within-subjects design by setting the layout as a between-subjects factor and others as within-subjects factors. Also, reflecting on the feedback on the readability of the graphs, the number of graphs was decreased from eight, sixteen, and thirty-two to six, twelve, and twenty-four for each of small, medium, and large datasets. Considering that most studies on the number of stocks for the substantial diversification suggest six to fifteen stocks, and some suggest over fifty or even more than 73 (Alexeev, Tapon, et al., 2014), these numbers of graphs for each size of dataset looked reasonable. Furthermore, reflecting on the feedback that there are too many points on each graph, making not only less readable but also hard to figure out the exact point to compare in the maximum/minimum task or the slope task, the number of points were decreased to 1/8 by leaving every first one from every eight points.

The second pilot study was conducted with fifteen participants on Prolific. Half the participants were randomly assigned to use line charts, and the other half were to use horizon graphs. The second pilot results suggested a very low accuracy of the maximum/minimum and

slope tasks with horizon graphs. Because the low accuracy of the maximum/minimum task with horizon graphs was not expected, the possibility that the participants might have been confused was considered, and the solutions to make the task clearer was applied. Considering that the red color used to represent the positive area, over the baseline, was also used to signify something dangerous and negative, the color scheme for horizon graphs was changed from the red-blue to the green-red, so that the positive, over the baseline area became green, and the negative, below the baseline area, became red. Also, to make use of the orientation as an additional channel to visualize, the mode of horizon graphs was changed from the mirror mode to the offset mode (See 3.4). Though it was indicated that the use of the offset mode didn't have an impact on the performance of using horizon graphs (Heer et al., 2009), it was expected that it would help the participants be less confused about which area was positive and negative. Also, the use of the offset mode was expected to help participants in the slope task because, with the offset mode horizon graph, they didn't need to flip the negative area to calculate the slope.

The third pilot study was conducted with twenty participants on Prolific. The results of the third pilot suggested that the use of the green-red color scheme and the offset mode worked for improving the accuracy of the horizon graphs, though this impact was not tested statistically. But the accuracy of the slope task was still very low for both line charts and horizon graphs, and specifically, that of the trials with the long span, requiring to consider the whole span of each graph, was significantly low. Therefore, asking participants to do the slope task with the long span was regarded meaningless, and the length of the span was changed from 1, 1/2, and 1/4 of the whole span of each graph to 1/2, 1/3, and 1/4.

After these three rounds of the pilot studies, the sample size of the primary study was determined and the primary study was conducted with 196 participants on Prolific.

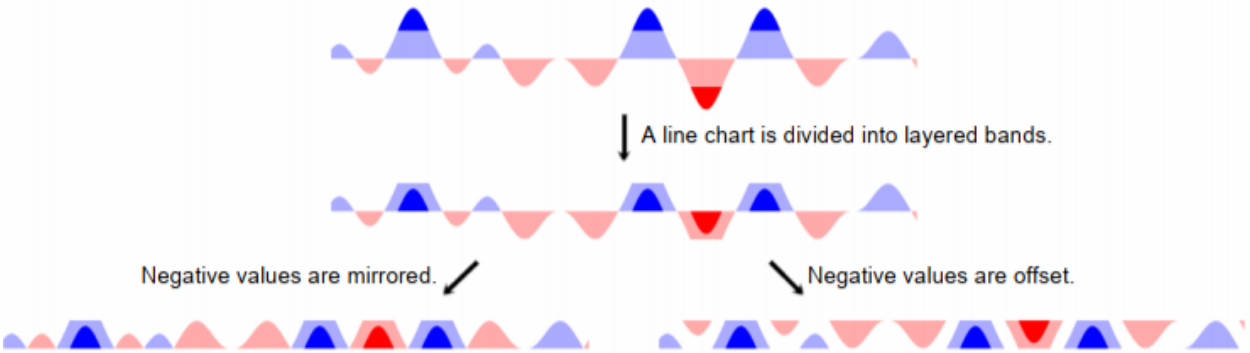


Figure 3.4. Construction of HG (adapted from Heer et al. (2009)

CHAPTER 4. RESULTS

In the first section of this chapter, the primary study results are overviewed in three parts: completion time, accuracy, and subjective responses. Each part provides the mean and standard deviation for each condition. From the data, the outliers under $mean - 1.5 * IQR$ or over $mean + 1.5 * IQR$ for each metric were removed (Tukey, 1977). And in the second section, the results are analyzed centered on four pairs of conditions corresponding to the four hypotheses for each.

4.1 Overview

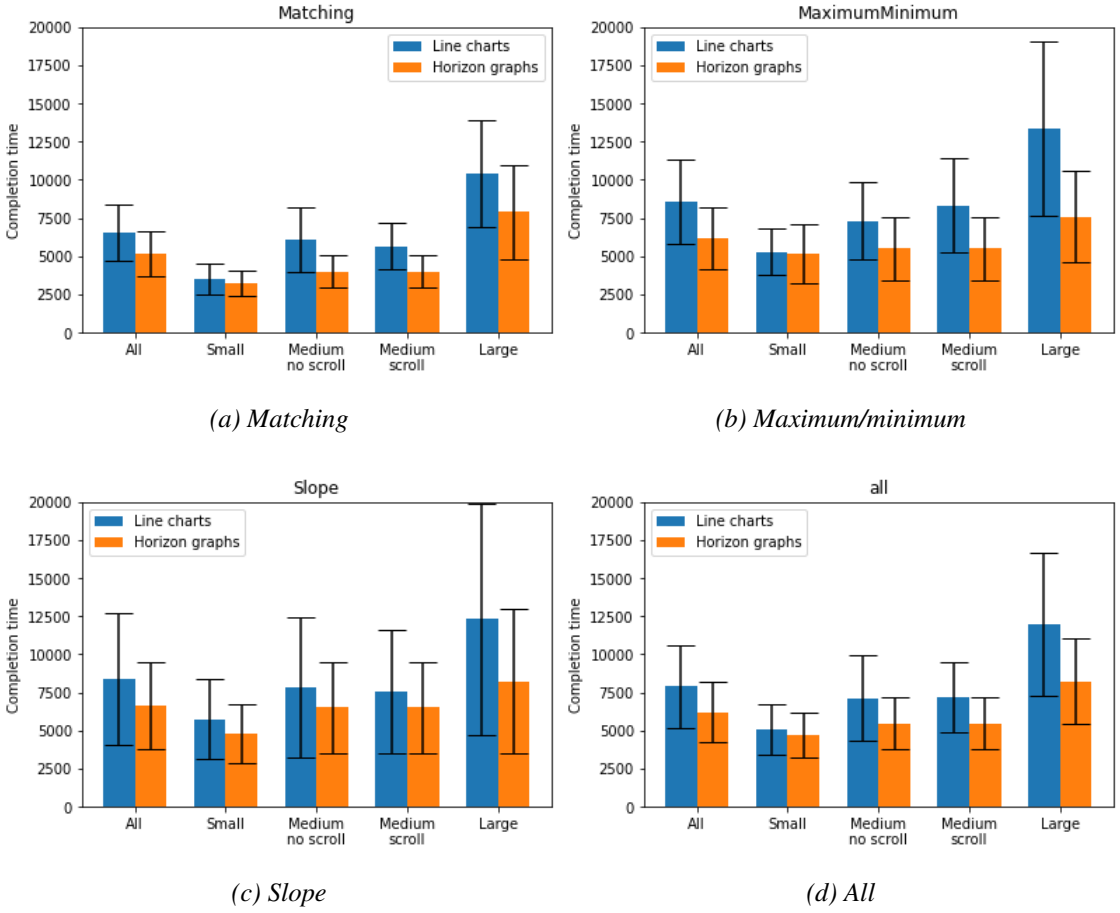
4.1.1 Completion Time

Figure 4.1 shows the mean and standard deviation of the completion times with columns as tasks. The blue bars represent the completion times of LC, and the orange bars represent those of HG. The unit of completion time on the y-axis is *ms*. The standard deviation of each condition is represented using black lines. Since scrolling was not used for the medium size data with HG, the bars for HG in the case of "Medium/no scroll" and "Medium/scroll" are the same. Along with the mean and standard deviation of the completion times, the statistics of the T-test, p-value and *d*

Table 4.1. *Pairs of Conditions*

Pairs	Size of dataset	Vertical scrolling interaction	
		LC	HG
Pair 1	Small	Not used	Not used
Pair 2	Medium	Not used	Not used
Pair 3		Used	Not used
Pair 4	Large	Used	Used

as a measure of the effect size, where 0.2 is a small size, 0.5 is a medium size, and 0.8 is a large size (Cohen, 2013), are reported in the Table 4.2.



(a) Matching *(b) Maximum/minimum*

(c) Slope *(d) All*

Figure 4.1. Completion Times for Each Task

For all the tasks, the completion times increased as the size of datasets increased. Also, as the size of datasets increase, the conditions with LC showed a more rapid increase in the completion time than HG. Standard deviations also more rapidly increase with LC than HG. For all tasks, HG took less time than LC. *Matching* task took less time with HG, which took 5.16 seconds on average for all conditions than LC, which took 6.55 seconds on average for all conditions (See Fig 4.1). *Maximum/minimum* task with LC took 8.61 seconds on average for all conditions, which was larger than 6.17 seconds of HG. *Slope* task with LC took 8.40 seconds on average for all conditions, which was larger than 6.62 seconds of HG.

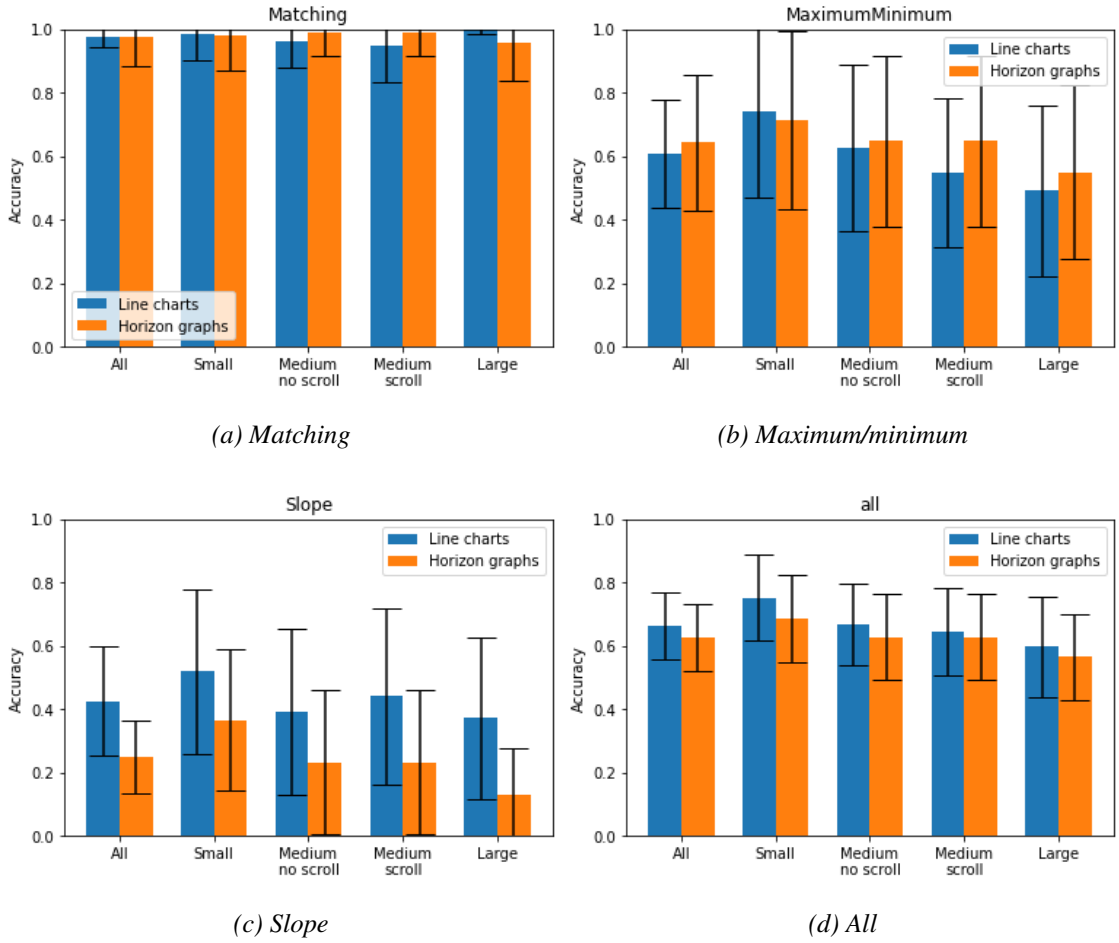
Table 4.2. *Effects of Layout on Completion Time*

Task	Pair	T	p	*d*
All	All	4.890	***	0.723
	Pair 1 (small size data)	1.396		0.207
	Pair 2 (medium size data, no scroll)	4.822	***	0.715
	Pair 3 (medium size data, scroll in LC)	5.644	***	0.844
	Pair 4 (large size data, scroll in both)	6.586	***	0.968
Matching	All	5.672	***	0.832
	Pair 1 (small size data)	2.157	*	0.323
	Pair 2 (medium size data, no scroll)	8.167	***	1.212
	Pair 3 (medium size data, scroll in LC)	8.295	***	1.251
	Pair 4 (large size data, scroll in both)	5.275	***	0.769
Maximum/minimum	All	6.762	***	1.003
	Pair 1 (small size data)	0.471		0.070
	Pair 2 (medium size data, no scroll)	5.392	***	0.789
	Pair 3 (medium size data, scroll in LC)	7.457	***	1.094
	Pair 4 (large size data, scroll in both)	8.548	***	1.264
Slope	All	3.245	**	0.482
	Pair 1 (small size data)	2.770	**	0.418
	Pair 2 (medium size data, no scroll)	2.319	*	0.345
	Pair 3 (medium size data, scroll in LC)	1.928		0.288
	Pair 4 (large size data, scroll in both)	4.361	***	0.641

*: $p < .05$, **: $p < .01$,***: $p < .001$

4.1.2 Accuracy

Figure 4.2 shows the accuracy and standard deviation of the accuracy with the column as tasks. As above, the blue bars represent LC, whereas the orange bars represent HG. Also, as above, scrolling was not used for the medium size data with HG, and the bars for HG in the case of "Medium/no scroll" and "Medium/scroll" are the same one with the medium size dataset. Along with the means and standard deviations, the statistics of T-test, p-value, and d as a measure of the effect sizes, where 0.2 is a small size, 0.5 is a medium size, and 0.8 is a large size (Cohen, 2013), are reported in the Table 4.3. Figure 4.2 shows that overall accuracy decreases as the size of the dataset increases. Also, the accuracy of the conditions with the *Matching* task was the highest, and the *Maximum/minimum* task was the next, and that of the *Slope* task was the lowest. It is notable that the accuracy of the *Slope* task with HG more rapidly decreased as the size of the dataset increased. Considering it with the high standard deviation of the *Slope* task conditions with HG, it is likely that the task was too difficult to find a clear impact of the increase of the size of the dataset.



(a) Matching

(b) Maximum/minimum

(c) Slope

(d) All

Figure 4.2. Accuracy for Each Task

4.1.3 Subjective Responses

The means and standard deviations of the subjective responses about the completion time (see Figure 4.3) and accuracy (see Figure 4.4) are reported. Also, the T-test statistics, p-values, and *d* as a measure of the effect size, where 0.2 is a small size, 0.5 is a medium size, and 0.8 is a large size(Cohen, 2013), are reported.

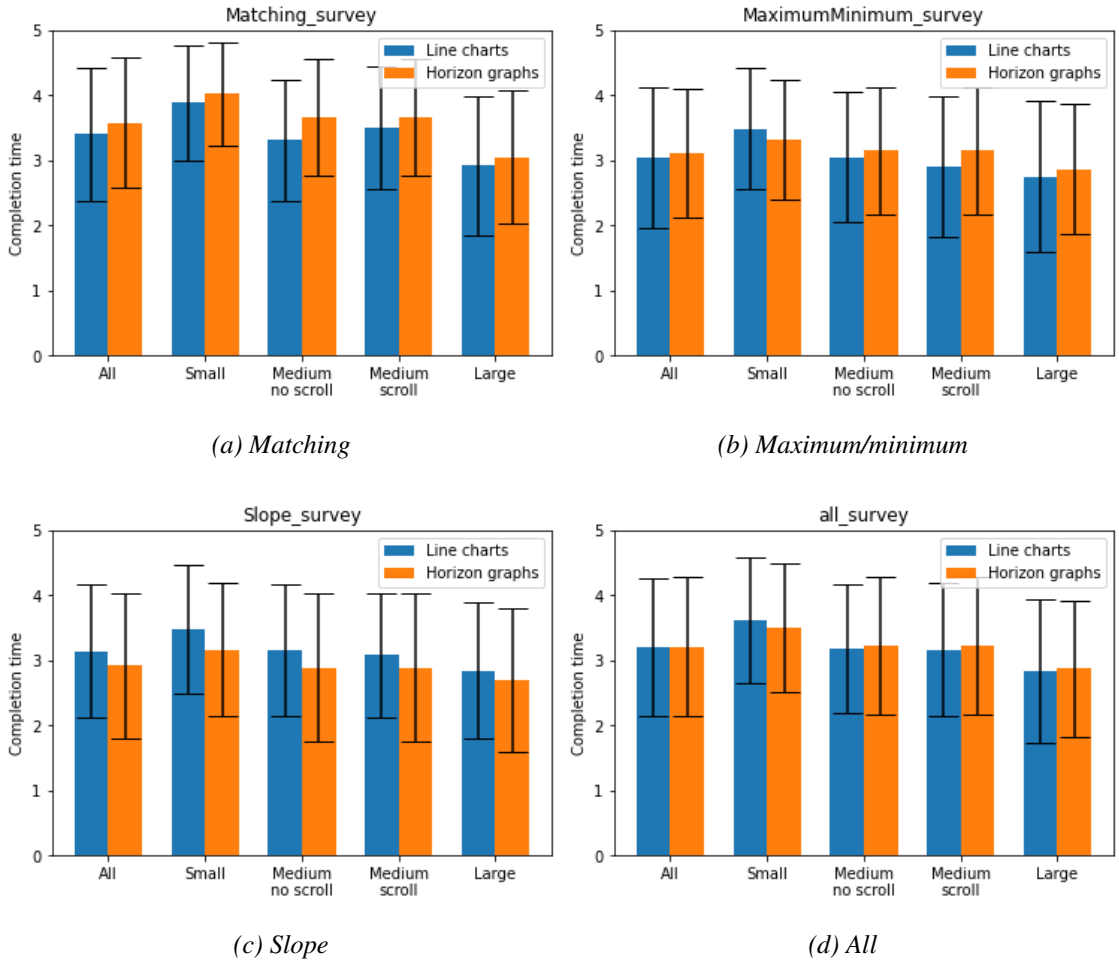
Overall, the trends in the subjective responses follow those of the completion time and accuracy, as the size of the datasets increases and the task proceeded from the *Matching* to *Maximum/minimum* and *Slope* task. But about the *Slope* task, though the participants did trials significantly faster, the survey results of the *Slope* task show that the participants who did the

Table 4.3. *Effects of Layout on Accuracy*

Task	Pair	T	p	*d*
All	All	2.229	*	0.353
	Pair 1 (small size data)	2.120	*	0.479
	Pair 2 (medium size data, no scroll)	2.101	*	0.306
	Pair 3 (medium size data, scroll in LC)	0.985		0.142
	Pair 4 (large size data, scroll in both)	1.509		0.217
Matching	All	0.059		0.009
	Pair 1 (small size data)	0.237		0.034
	Pair 2 (medium size data, no scroll)	-2.489	*	-0.365
	Pair 3 (medium size data, scroll in LC)	-2.759	**	-0.397
	Pair 4 (large size data, scroll in both)	2.999	**	0.465
Maximum/minimum	All	-1.270		-0.184
	Pair 1 (small size data)	0.739		0.106
	Pair 2 (medium size data, no scroll)	-0.598		-0.086
	Pair 3 (medium size data, scroll in LC)	-2.696	**	-0.388
	Pair 4 (large size data, scroll in both)	-1.483		-0.214
Slope	All	7.968	***	1.183
	Pair 1 (small size data)	4.387	***	0.632
	Pair 2 (medium size data, no scroll)	4.546	***	0.656
	Pair 3 (medium size data, scroll in LC)	5.662	***	0.817
	Pair 4 (large size data, scroll in both)	7.749	***	1.147

*: $p < .05$, **: $p < .01$,***: $p < .001$

tasks with the HG were less confident than the ones who did with the LC. Chances are that they were less confident in their completion time because, as the low accuracy of the *Slope* task conditions with HG show, the trials were more difficult than the ones they did in other conditions.



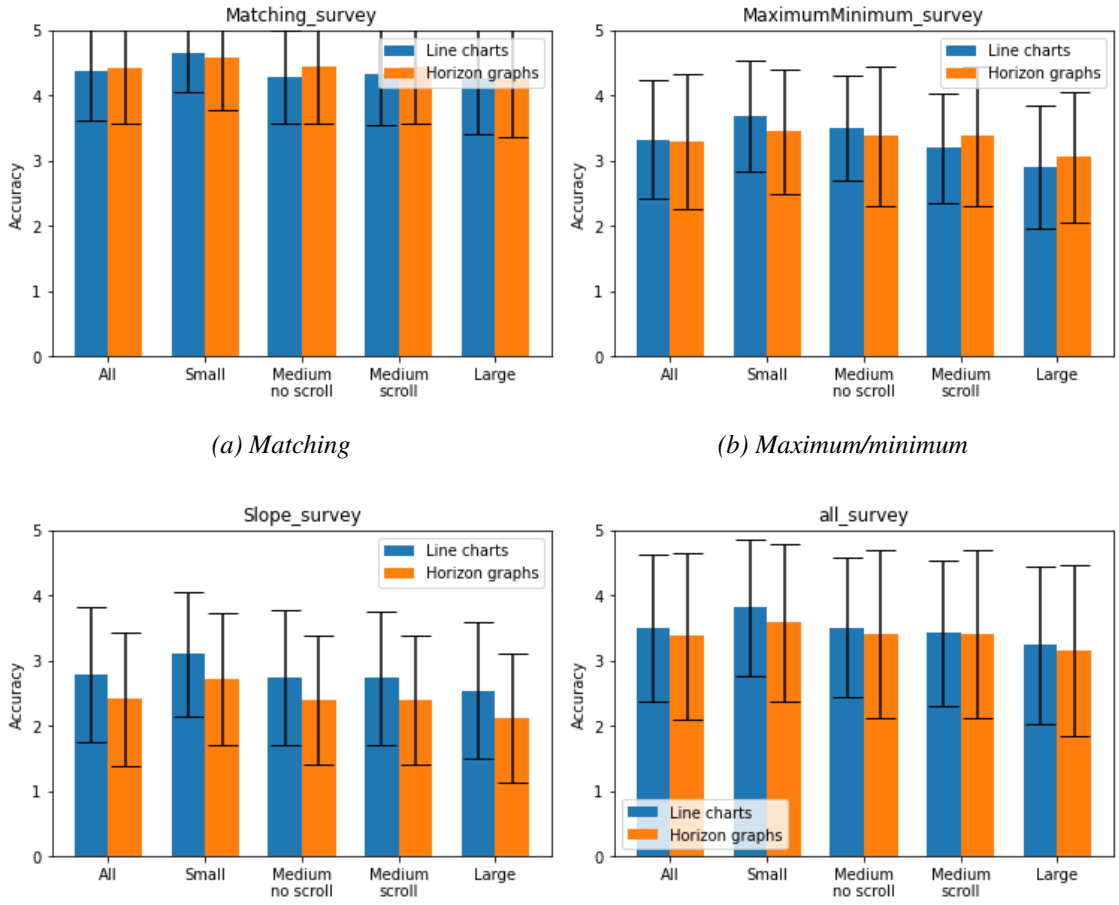
(a) Matching

(b) Maximum/minimum

(c) Slope

(d) All

Figure 4.3. Survey Results on Completion Time



(a) Matching *(b) Maximum/minimum*

(c) Slope *(d) All*

Figure 4.4. Survey Results on Accuracy

4.2 Result Analyses

In this section, the results of the hypotheses test with the independent t-test are reported. The independent t-test of LC and HG was conducted for the following four pairs of different conditions. By pair 1 and pair 2, the basic graphical perception of two different visual representations was compared. Also, with these pairs, the effect of the low height of LC on readability and that of using HG were compared. Pair 1 was a baseline for comparing two layouts because representing small data using LC does not lose readability in the given vertical space.

Table 4.4. *Effects of Layout on the Subjective Response for Completion Time*

Task	Pair	T	p	*d*
All	All	-0.111		-0.030
	Pair 1 (small size data)	1.348		0.127
	Pair 2 (medium size data, no scroll)	-0.698		-0.101
	Pair 3 (medium size data, scroll in LC)	-0.788		-0.104
	Pair 4 (large size data, scroll in both)	-0.392		-0.061
Matching	All	-2.243	*	-0.249
	Pair 1 (small size data)	-1.170		-0.188
	Pair 2 (medium size data, no scroll)	-2.650	**	-0.415
	Pair 3 (medium size data, scroll in LC)	-1.191		-0.208
	Pair 4 (large size data, scroll in both)	-0.875		-0.138
Maximum/minimum	All	-0.777		-0.093
	Pair 1 (small size data)	1.205		0.161
	Pair 2 (medium size data, no scroll)	-0.669		-0.121
	Pair 3 (medium size data, scroll in LC)	-1.591		-0.239
	Pair 4 (large size data, scroll in both)	-0.765		-0.135
Slope	All	2.757	**	0.383
	Pair 1 (small size data)	2.170	*	0.303
	Pair 2 (medium size data, no scroll)	1.770		0.212
	Pair 3 (medium size data, scroll in LC)	1.290		0.257
	Pair 4 (large size data, scroll in both)	0.953		0.554

*: $p < .05$, **: $p < .01$,***: $p < .001$

Table 4.5. *Effects of Layout on the Subjective Response for Accuracy*

Task	Pair	T	p	*d*
All	All	2.128	*	0.169
	Pair 1 (small size data)	2.485	*	0.349
	Pair 2 (medium size data, no scroll)	0.967		0.122
	Pair 3 (medium size data, scroll in LC)	0.066		-0.003
	Pair 4 (large size data, scroll in both)	0.854		0.103
Matching	All	-0.723		-0.089
	Pair 1 (small size data)	0.849		0.094
	Pair 2 (medium size data, no scroll)	-1.369		-0.249
	Pair 3 (medium size data, scroll in LC)	-0.978		-0.187
	Pair 4 (large size data, scroll in both)	-0.021		-0.006
Maximum/minimum	All	0.388		0.027
	Pair 1 (small size data)	1.824		0.253
	Pair 2 (medium size data, no scroll)	0.881		0.110
	Pair 3 (medium size data, scroll in LC)	-1.397		-0.212
	Pair 4 (large size data, scroll in both)	-1.021		-0.154
Slope	All	4.627	***	0.413
	Pair 1 (small size data)	2.756	**	0.381
	Pair 2 (medium size data, no scroll)	2.327	*	0.318
	Pair 3 (medium size data, scroll in LC)	2.290	*	0.330
	Pair 4 (large size data, scroll in both)	2.900	**	0.397

*: $p < .05$, **: $p < .01$,***: $p < .001$

Table 4.6. *Hypotheses Test Results*

Hypothesis	Test result
H1: Using the small dataset, without any scrolling, participants will perform tasks faster with LC.	Not supported
H2: Using the small dataset, without any scrolling, there will be no significant difference in task accuracy between LC and HG.	Not supported
H3: Using the medium dataset without any scrolling, HG will outperform LC, with higher accuracy and less completion time.	Partially supported
H4: Using the medium dataset, HG will be more effective than LC when vertical scrolling interaction is used in LC and is not used in HG.	Supported
H5: Using the large dataset, HG will be more effective than LC when vertical scrolling interaction is used in both cases.	Supported

With pair 3 with the medium dataset and vertical scrolling only in LC, the effect of interaction, and that of consulting memory and using HG to performance were compared. In this pair of conditions, the virtual resolution of the display of LC was twice that of HG, and the virtual resolution of each horizon graph was equal to that of an LC because it used half the space with a 2-band offset.

Lastly, with pair 4 with the large dataset and vertical scrolling in both LC and HG, the effect of using HG for consulting to memory was examined. In this case, to make HG involve vertical scrolling, the virtual resolution of the display in the case of HG was twice of a single display. It used two pages of display. And the virtual resolution of the display in the case of LC was twice that of HG to make the virtual display, with four pages (See Table 3.2). In this condition, tasks were comparing line graphs or HG displayed on the different pages requiring scroll.

For all the results of the t-test for these pairs, refer to Table 4.2 and 4.3.

4.2.1 Pair 1: with the Small Dataset

With pair 1, the performances of LC and HG with the small dataset, six graphs are compared. In this pair, by the definition of the small dataset in this study, scrolling was not used for either of LC and HG. In this pair, the heights of LC and HG were the same. However, since LC and HG were using the same vertical space for the same number of graphs, the virtual resolution of HG was quadruple that of LC because the height of each of the positive and negative span was twice that of HG as layered. There was a statistically significant difference in the completion time between LC and HG in the results of all of the three task trials. HG was significantly faster than LC. For the detailed results, refer to the Table 4.2.

There was no statistically significant difference in the accuracy between LC and HG in the *Matching* and *Maximum/minimum* task conditions. However, in the *Slope* conditions, HG was significantly less accurate than LC (See Table 4.3).

Based on this result, H1, using the small dataset, without any scrolling, participants will perform tasks faster with LC, is not supported. And H2, using the small dataset, without any scrolling, there will be no significant difference in task accuracy between LC and HG is not

supported. Overall, the LC was more accurate than HG, especially because of the difference in the accuracy of the *Slope* task. Unlike the initial expectation that LC will outperform HG when there is no clear advantage of using HG for saving vertical space, HG outperformed LC except for the *Slope* task where no significant difference existed.

4.2.2 Pair 2: with the Medium Dataset, without Scrolling for Both LC and HG

With pair 2, the performances of LC and HG with the medium dataset, twelve graphs were compared. In this pair, scrolling was not used in both LC and HG. To provide twelve graphs without scrolling, the height of each LC in this pair was half that of the small dataset. In this case, the heights of both LC and HG were the same, but the virtual resolution of HG was quadruple that of LC.

In these conditions in all the tasks, there were statistically significant differences in the completion time between LC and HG. HG was significantly faster than LC for all the tasks. Also, considering the overall effect size of this pair with pair 1, the advantage of using HG in terms of completion time was more clear. The d of completion time in pair 1 was 0.207, which was small (Cohen, 2013), whereas that in pair 2 was 0.715, which was between the medium size effect, 0.4, and the large size effect, 0.8 (Cohen, 2013).

In terms of accuracy, LC overall outperformed HG($p < .05, d = .232$). More specifically, though HG was more accurate in the *Matching* ($p < .05, d = -0.365$) and more or as much accurate as LC in he *Maximum/minimum* task conditions ($p > .05, d = -0.086$), HG was far significantly more accurate in the *Slope* conditions ($p < .001, d = 0.656$).

Based on the results above, H3 is partially supported because overall, HG is faster but less accurate than LC.

4.2.3 Pair 3: with the Medium Dataset, Scrolling Only in LC

With pair 3, the performances of LC and HG with the medium dataset, twelve graphs were compared. In this pair, scrolling was used in LC but not in HG. In the case of LC, since twelve graphs were represented using two pages, the height of a graph in LC conditions was the same as

that of LC with the small dataset. Since the height of each HG was half that of each LC, the virtual resolution of HG was double that of LC.

In this condition, in all the tasks, there were statistically significant differences in the completion time between LC and HG. HG was significantly faster than LC. Comparing the effect sizes of this pair with those of pair 2, not using scroll, the effect sizes of pair 3 is larger than those of pair 2 except for the *Slope* task condition, the difference of overall was .129, in the *Matching* task conditions, .039, in the *Maximum/minimum* conditions, .305, and in the *Slope* conditions, -.063. In terms of accuracy, HG outperformed in the *Matching* and *Maximum/minimum* conditions, but underperformed in the *Slope* conditions.

Based on these results, H4 is supported because overall, HG was faster than LC and as accurate as LC.

4.2.4 Pair 4: with the Large Dataset, Scrolling in Both LC and HG

With pair 4, the performances of LC and HG with the large dataset, twenty-four graphs were compared. In this pair, scrolling was used in both LC and HG. In the case of LC, since twenty-four graphs were represented using four pages, six graphs per page, the height of a graph in LC conditions were the same as that of LC with the small dataset. HG was represented using two pages. Therefore the height for a graph in HG conditions was the same as that of HG with the medium dataset. Since the height of a graph of LC was double that of a graph of HG, the virtual resolution of HG was double that of LC.

In this condition, in all the task conditions, there were statistically significant differences in the completion time between LC and HG. HG was significantly faster than LC. In terms of accuracy, overall, there was no statistically significant difference between HG and LC. But LC was more accurate in the *Matching* and the *Slope* task conditions and HG was more accurate in the *Maximum/minimum* task condition.

Based on these results, H5 is supported because overall, HG is faster and as accurate as LC.

CHAPTER 5. DISCUSSION

Based on the results in the previous chapter, the implication of the results is discussed with consideration of the research questions. The design implications of the results, limitations of this study, and future work are also discussed.

5.1 Understanding the Perception of HG and Participants' Strategy using HG

The results of the conditions with the small dataset were different from the initial expectations. Initially, LC was expected to be better for the tasks than HG with the small dataset because, with the small dataset, HG would require additional cognitive load than line charts for mentally unstacking, interpreting layering and mirroring or offset, and virtually implementing the shape of line charts in mind. Also, with the small size dataset, it was assumed that LC provides enough readability so that HG was expected not to have an advantage of better showing variations than LC. However, HG outperformed or performed as much as LC with the small dataset, and one of the potential explanation is using the preattentive processing (Healey & Enns, 2011; Treisman, 1985). The preattentive processing is the detection of a limited set of visual features in less than 200-250 milliseconds (Healey & Enns, 2011). The preattentive visual features include hue (Bauer, Jolicoeur, & Cowan, 1997; D'Zmura, 1991; Healey & Enns, 1999; KAWAI, Uchikawa, & Ujike, 1995; Nagy & Sanchez, 1990), orientation (Julesz & Bergen, 1983; Sagi & Julesz, 1984; Weigle et al., 2000; Wolfe, Friedman-Hill, Stewart, & O'Connell, 1992), length (Sagi & Julesz, 1985; Treisman & Gormican, 1988), size (Healey & Enns, 1999; Treisman & Gelade, 1980), curvature (Treisman & Gormican, 1988), etc. (Healey & Enns, 2011). Considering that the use of hue for the positive/negative values is the clearest difference between HG and LC, hue might have an impact on the perception of HG.

Another way of understanding the perception of HG is to directly ask the strategy they used to complete the tasks with HG. If the approach using the theories about the preattentive

processing is low-level, asking about the strategy used is more of a higher level since the participants need to actively make a decision on which strategy they would use and find their own strategy as they become more familiar with using HG. As future work, the strategies used by the participants should be examined by interview or survey.

5.2 Implication for Design

As indicated in the first chapter, the purpose of this study is to examine whether using HG for visualizing time series data is a viable design choice for mobile phones when compared to LC. Considering this goal, the research question of this study was *What are the performance differences between HG and LC on the mobile phone display?* This research question was divided into two subquestions, one about the differences in completion time and another about the differences in accuracy.

In terms of completion time, HG was faster than LC, even in the case with the small size data where LC was expected to outperform. From the results with different conditions, the advantage of using HG for shorter completion time is clear. This result is also meaningful, considering subjective responses. Even though it was expected that participants would not be familiar with using HG than LC, the results suggest that the participants who used HG performed the tasks faster than those who used LC and also were more or as much confident in their completion time as those who used LC (See Table 4.4).

In terms of accuracy, though overall results suggest that LC was more accurate, especially the accuracy should be considered with the detailed results of each task since the *Matching* and *Maximum/minimum* tasks and the *Slope* task show strong contrast in the results. As suggested in Figure 4.2 and Table 4.3, in the *Matching* and *Maximum/minimum* task conditions, HG was more accurate or as much accurate as LC. And there was no significant difference in the results of the overall subjective responses. But in the *Slope* task conditions, LC was significantly accurate in all the pairs. However, the accuracy of HG in the *Slope* task conditions were about 40% with the small dataset, slightly over 20% with the medium dataset, and lower than 20% with the large dataset, which is too low to use in the field. Though the accuracy of LC in the *Slope* task condition is also lower than that of other tasks, that of HG rapidly decreased in this condition.

And the subjective responses to the *Slope* task suggest that the participants were significantly more confident in using LC for the task.

Also, considering the use of space, HG showed better completion time while using less space than LC in pairs 2, 3, and 4. For the pair 2 and 3 in *Matching* and *Maximum/minimum* task conditions, HG was more accurate than LC or as much accurate as LC. But for pair 4, LC was significantly more accurate than HG in the *Matching* and *Slope* tasks, and as much accurate in the *Maximum/minimum* task.

Therefore, HG can be potentially considered as a design choice over LC in the context where the *Slope* task is not expected to be important, in that it can save time without losing much accuracy. However, it should be considered when scrolling is used with HG because the accuracy highly depends on the task. More specifically, considering that this study is about the low-level primitive graphical perception study which used no visual aids used in real-world cases, such as ticks, values, and other information about data that can help participants, it is concluded that the experiment results show the high potential of HG as an alternative to LC. With more visual aids that can help improve the advantage of using HG and recover the disadvantages of using HG in cases such as where the *Slope* task is expected, HG can be a good alternative to LC. How to harness the potential advantage of using HG can be a meaningful next step.

5.3 Impact of Using Scrolling Interaction and the Different Size of Dataset

This study is mainly focused on examining the performance difference between HG and LC. However, by comparing the absolute effect sizes of pair 2 and pair 3, the impact of scrolling and small size on the performance using LC can be inferred.

In terms of completion time, considering the effect sizes of pair 2 and pair 3, except for the *Slope* task conditions, the effect sizes of pair 3 were larger than that of pair 2 (See Table 4.2. When aggregated, the *d* of pair 2 was 0.715, pair 3, 0.844, with the *Matching* task, that of pair 2 was 1.212, pair 3, 1.251, and with the *Maximum/minimum* task, that of pair 2 was 0.789, and pair 3, 1.094). From these results, it can be inferred that the use of scrolling has a larger impact on the completion time of using LC than the use of smaller graphs.

In terms of accuracy, considering the absolute effect sizes of pair 2 and pair 3, except for the *Slope* task conditions, the absolute effect sizes of pair 3 were large than those of pair 2 (See Table 4.3. With the *Matching* task, the absolute *d* of pair 2 was 0.365, pair 3, 0.397, and with the *Maximum/minimum* task, that of pair 2 was 0.086 and pair 3, 0.388). From these results, it can be inferred that the use of scrolling has a larger impact on the accuracy of using LC than the use of smaller graphs.

Except for the *Slope* task conditions, the results above suggest that scrolling has a larger impact on the performance than using the graphs with smaller size.

5.4 Optimal Conditions for Using HG

In this study, it was found that HG outperforms LC, and the effect of using HG was larger, especially when the number of graphs on a display increased. Exploring more tasks that can fully take advantage of HG would be a meaningful next step.

Also, though this study initially followed the guideline suggested by (Heer et al., 2009) to choose the features of HG, including the number of bands, mode, and color scheme, running the study for checking if the optimal conditions suggested by (Heer et al., 2009) is also optimal for HG on mobile phones will be a meaningful contribution for using HG on mobile phones. Specifically, in the pilot study, it was found that the accuracy of HG increased when the mode of HG was changed from mirror to offset. Because it was only the result of the pilot study with a small sample size, fifteen and twenty, it was not examined enough. As a part of finding the optimal conditions for using HG on mobile phones, this can be checked as well.

Also, this study set the number of graphs on a display referring to the real use cases and reflecting on the feedback from the participants of the pilot study. But finding the optimal number of graphs on a display would be useful.

5.5 Comparison to the Previous Studies

This study is the first study that examined the performance of using HG on a mobile display. Also, this study explicitly and actively used a priori power analysis to decide the exact

sample size to make sure the statistical results with enough statistical power. By following this process, this study recruited 196 participants for the primary study, which is far more than previous studies, which mostly recruited less than or equal to 20 participants (Heer et al., 2009; Jabbari et al., 2018a, 2018b; Javed et al., 2010; Perin et al., 2013).

In terms of the results, it is notable that unlike what Javed et al. (2010) reported that the visualization type did not have a significant effect on accuracy, this study reports that the layout (LC or HG) had an impact on the accuracy overall and particularly in the *Slope* task conditions (See Table 4.3). As suggested in Perin et al. (2013), it is likely that the use of the real-world dataset with large and small scale variation (Perin et al., 2013) or the difference in the size of the display were related to this result.

5.6 Limitations

Since the experiment of this study was conducted using a crowdsourcing platform, there were inevitable tradeoffs between ecological validity and control over the participants. Though about two hundred participants were recruited shortly and the limitation of the control over them is advantageous for ecological validity in that it realizes the context of use in everyday life of participants using mobile phones, participating in an experiment, and doing the tasks for 25 minutes were not a very realistic situation. However, by designing and implementing the features for quality control and by checking the error and completion time of participants together, quality data could be collected. Since Prolific supported to reject the submissions that did not pass the attention checks, that with too low accuracy, or that did not follow the instruction appropriately, the submissions with the low quality could be excluded without losing the number of participants.

Also, how the test program is viewed cannot be fully controlled. First, depending on the resolution and the status of the display of devices participants use, the size and color of the graphs can be different. When a participant uses the color filters, such as the blue light filter and the filters for color blindness, the graphs can be seen differently. A researcher cannot fully control the settings of the Operation System level. Also, though how the test program is viewed on a mobile browser can be controlled and it did not allow the participants to use the zoom interaction, technically, if the zooming is set by the Operating System level, this cannot be controlled by the

researcher. Since most participants were in the 20s or 30s ($mean = 25.58, std = 6.66$), most of them are unlikely to use the zooming or magnifying feature on the Operating System level for helping their weak eyesight.

As indicated above, this study is a low-level primitive graphical perception study, which compared the pure perception of HG and LC without a very specific context of use. Though the study aimed at high ecological validity of the results by running an online crowdsourcing experiment using the real-world data on mobile display settings, there were no visual aids, such as ticks, specific values on the axes, names of the stock items, which are common in the real-world context. The tasks were more primitive ones, and the questions for each task was more decontextualized and abstract than the queries in the real-world use cases. Therefore, the results of this study are not directly connected to the practical design guidelines that can be directly used in the field. It is rather suggesting the potential of HG as an alternative to LC when properly used.

Lastly, the color scheme used for HG in this study was found not to support the participants with red-green color blindness. Since the positive and negative value ranges were redundantly encoded with color and the position of the starting point of a range, positive from the bottom and negative from the top, even the people with red-green color blindness can read the graph. However, it might have had an impact on the performance of those with red-green color blindness.

CHAPTER 6. CONCLUSION

In this study, the viability of HG as an alternative to LC was examined by running the crowdsourcing experiments with different conditions by the size of the dataset and vertical scrolling. The results of the experiment suggest that HG is a viable design and can be an excellent alternative to LC to represent a large number of multiple time series data. In the experiment, in most cases, HG outperformed LC, except for the *Slope* task, which was the most challenging task even with LC and was turned out that HG was not appropriate for. Even in the case where LC was initially expected to outperform HG, HG was more or as accurate as LC while taking much less time. In addition to supporting to use of HG as an alternative to LC to visualize multiple time-series data, these results leave further works to understand why HG outperforms LC. This could be approached in terms of low-level, using the preattentive processing theories (Healey & Enns, 2011), or high-level by asking participants about their strategy. Since HG has multiple visual properties, including hue, the brightness of color, shape of the area, and different starting points to represent positive/negative values, there can be various strategies to complete the tasks fast and accurately.

REFERENCES

Aigner, W., Miksch, S., Müller, W., Schumann, H., & Tominski, C. (2007). Visualizing time-oriented data-A systematic view. *Computers and Graphics (Pergamon)*, *31*(3), 401–409. doi: 10.1016/j.cag.2007.01.030

Alexeev, V., Tapon, F., et al. (2014). The number of stocks in your portfolio should be larger than you think: Diversification evidence from five developed markets. *J. Invest. Strateg*, *4*, 43–82.

Amir, O., Rand, D. G., et al. (2012). Economic games on the internet: The effect of $1 stakes. *PloS one*, *7*(2).

Andrienko, N., & Andrienko, G. (2006). *Exploratory analysis of spatial and temporal data: A systematic approach*. Springer Scienc. doi: 10.1007/3-540-31190-4

Apple. (n.d.-a). *Activity*. Retrieved 2019-12-17, from `https://apps.apple.com/us/app/activity/id12082247953`

Apple. (n.d.-b). *iOS - Health - Apple*. Retrieved 2020-01-25, from `https://www.apple.com/lae/ios/health/`

Apple. (n.d.-c). *iOS - Stocks*. Retrieved 2020-03-02, from `https://apps.apple.com/us/app/stocks/id1069512882`

Apple. (n.d.-d). *Weather on the App Store*. Retrieved 2020-01-25, from `https://apps.apple.com/us/app/weather/id1069513131`

Archambault, D., Purchase, H., & Pinaud, B. (2011). Animation, small multiples, and the effect of mental map preservation in dynamic graphs. *IEEE Transactions on Visualization and Computer Graphics*, *17*(4), 539–552. doi: 10.1109/TVCG.2010.78

Badam, S. K. (2018). Towards a Unified Visualization Platform for Ubiquitous Analytics. In *EA '18: Extended Abstract Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. ACM.

Bauer, B., Jolicoeur, P., & Cowan, W. (1997). Visual search for colour targets that are or are not linearly separable from distractors. *Ophthalmic Literature*, *1*(50), 53.

Beattie, V., & Jones, M. J. (2002). The impact of graph slope on rate of change judgments in corporate reports. *Abacus*, *38*(2), 177–199. doi: 10.1111/1467-6281.00104

Bederson, B. B., Clamage, A., Czerwinski, M. P., & Robertson, G. G. (2004). DateLens: A fisheye calendar interface for PDAs. *ACM Transactions on Computer-Human Interaction*, *11*(1), 90–119. doi: 10.1145/1005261.1005268

Blascheck, T., Besançon, L., Bezerianos, A., Lee, B., & Isenberg, P. (2019). Glanceable visualization: Studies of data comparison performance on smartwatches. *IEEE Transactions on Visualization and Computer Graphics*, *25*(1), 630–640. doi: 10.1109/TVCG.2018.2865142

Borgo, R., Lee, B., Bach, B., Fabrikant, S., Jianu, R., Kerren, A., … Michelle, Z. (2017). Crowdsourcing for Information Visualization: Promises and Pitfalls. In D. Archambault, H. Purchase, & T. Hoßfeld (Eds.), *Evaluating in the crowd. crowdsourcing and human-centered experiments* (pp. 96–138).

Borgo, R., Micallef, L., Bach, B., McGee, F., Lee, B. (2018). Information Visualization Evaluation Using Crowdsourcing. *Computer Graphics Forum*, *37*(3), 573–595. doi: 10.1111/cgf.13444

Bostock, M., Ogievetsky, V., & Heer, J. (2011). D^3 data-driven documents. *IEEE transactions on visualization and computer graphics*, *17*(12), 2301–2309.

Boyandin, I., Bertini, E., & Lalanne, D. (2012). A Qualitative Study on the Exploration of Temporal Changes in Flow Maps with Animation and Small-Multiples. *Computer Graphics Forum*, *31*(3pt2), 1005–1014. doi: 10.1111/j.1467-8659.2012.03093.x

Brehmer, M., Lee, B., Isenberg, P., & Choe, E. K. (2019a). A Comparative Evaluation of Animation and Small Multiples for Trend Visualization on Mobile Phones. *IEEE Transactions on Visualization and Computer Graphics*, 1–1. doi: 10.1109/tvcg.2019.2934397

Brehmer, M., Lee, B., Isenberg, P., & Choe, E. K. (2019b). Visualizing Ranges over Time on Mobile Phones: A Task-Based Crowdsourced Evaluation. *IEEE Transactions on Visualization and Computer Graphics*, *25*(1), 619–629. doi: 10.1109/TVCG.2018.2865234

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5. doi: 10.1177/1745691610393980

Byron, L., & Wattenberg, M. (2008). Stacked graphs - Geometry & aesthetics. *IEEE Transactions on Visualization and Computer Graphics*, *14*(6), 1245–1252. doi: 10.1109/TVCG.2008.166

Card, S. K., Mackinlay, J. D., & Shneiderman, B. (1999). *Readings in information visualization: using vision to think*. Morgan Kaufmann.

Chittaro, L. (2006a). Visualization of patient data at different temporal granularities on mobile devices. *Proceedings of the Workshop on Advanced Visual Interfaces*, *2006*, 484–487. doi: 10.1145/1133265.1133364

Chittaro, L. (2006b). Visualizing Information on Mobile Devices. *North*, 40–45.

Clement, J. (2019). *Percentage of all global web pages served to mobile phones from 2009 to 2018.* Retrieved from `https://www.statista.com/statistics/241462/global-mobile-phone-website-traffic-share/`

Cleveland, W. S., & McGill, R. (1984). *Graphical Perception Theory Experimentation and Application, Jrnl Americal Stats Assoc, Vol 79, No 387 Sept 1984.pdf* (Vol. 79) (No. 387).

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Academic press.

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating amazon's mechanical turk as a tool for experimental behavioral research. *PloS one*, *8*(3).

Dalton, N., Katz, D., & Keynes, M. (2018). Visualizing Diabetes data in Mobile contexts. In *Ea '18: Extended abstract proceedings of the acm conference on human factors in computing systems (chi).* ACM.

D'Zmura, M. (1991). Color in visual search. *Vision research*, *31*(6), 951–966.

Federico, P., Hoffmann, S., Rind, A., Aigner, W., & Miksch, S. (2014). Qualizon graphs: Space-efficient time-series visualization with qualitative abstractions. In *Proceedings of the 2014 international working conference on advanced visual interfaces* (pp. 273–280).

Few, S. (2008). Time on the Horizon. *Visual Business Intelligence Newsletter*, 1–7. Retrieved from `http://www.perceptualedge.com/articles/visual{\_}business{\_}intelligence/time{\_}on{\_}the{\_}horizon.pdf`

Fitbit. (2015). *Fitbit Official Site for Activity Trackers & More.* Retrieved from `https://www.fitbit.com/us/homehttp://www.fitbit.com/{\#}i.1mxltfgou4fqdt`

FiveThirtyEight. (n.d.). Retrieved from `https://fivethirtyeight.com/`

Fuchs, J., Fischer, F., Mansmann, F., Bertini, E., & Isenberg, P. (2013). Evaluation of alternative glyph designs for time series data in a small multiple setting. *Conference on Human*

Factors in Computing Systems - Proceedings, 3237–3246. doi: 10.1145/2470654.2466443

Gadiraju, U., Kawase, R., Dietze, S., & Demartini, G. (2015). Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. *Conference on Human Factors in Computing Systems - Proceedings*, *2015-April*, 1631–1640. doi: 10.1145/2702123.2702443

Gogolou, A., Tsandilas, T., Palpanas, T., & Bezerianos, A. (2019). Comparing Similarity Perception in Time Series Visualizations. *IEEE Transactions on Visualization and Computer Graphics*, *25*(1), 523–533. doi: 10.1109/TVCG.2018.2865077

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013, jul). Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples. *Journal of Behavioral Decision Making*, *26*(3), 213–224. doi: 10.1002/bdm.1753

Google LLC. (n.d.). *Google Maps*. Retrieved 2020-01-25, from `https://apps.apple.com/us/app/google-maps-transit-food/id585027354`

Guan, Z., & Cutrell, E. (2007). An eye tracking study of the effect of target rank on Web search. *Conference on Human Factors in Computing Systems - Proceedings*, 417–420. doi: 10.1145/1240624.1240691

Healey, C. G., & Enns, J. (2011). Attention and visual memory in visualization and computer graphics. *IEEE transactions on visualization and computer graphics*, *18*(7), 1170–1188.

Healey, C. G., & Enns, J. T. (1999). Large datasets at a glance: Combining textures and colors in scientific visualization. *IEEE transactions on visualization and computer graphics*, *5*(2), 145–167.

Heer, J., & Bostock, M. (2010). Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design. In *Proceedings of the 28th annual chi conference on human factors in computing systems* (pp. 203–212). doi: 10.1145/1753326.1753357

Heer, J., Kong, N., & Agrawala, M. (2009). Sizing the horizon: The effects of chart size and layering on the graphical perception of time series visualizations. *Conference on Human Factors in Computing Systems - Proceedings*. doi: 10.1145/1518701.1518897

Hochheiser, H., & Shneiderman, B. (2004). Dynamic Query Tools for Time Series Data Sets: Timebox Widgets for Interactive Exploration. *Information Visualization*, *3*(1), 1–18. doi: 10.1057/palgrave.ivs.9500061

Horak, T., & Dachselt, R. (2018). Hierarchical Graphs on Mobile Devices : A Lane-based Approach. In *Ea '18: Extended abstract proceedings of the acm conference on human factors in computing systems (chi).* ACM.

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental economics*, *14*(3), 399–425.

Howe, J. (2006). *Crowdsourcing: Crowdsourcing: A Definition.* Retrieved 2020-01-27, from `https://crowdsourcing.typepad.com/cs/2006/06/crowdsourcing{\_}a.html`

Huang, W., Eades, P., & Hong, S. H. (2008). Beyond time and error: A cognitive approach to the evaluation of graph drawings. *Proceedings of the 2008 Conference on BEyond Time and Errors: Novel EvaLuation Methods for Information Visualization 2008, BELIV'08*, 1–8. doi: 10.1145/1377966.1377970

Huang, W., Eades, P., & Hong, S. H. (2009). Measuring effectiveness of graph visualizations: A cognitive load perspective. *Information Visualization*, *8*(3), 139–152. doi: 10.1057/ivs.2009.10

Jabbari, A., Blanch, R., & Dupuy-Chessa, S. (2018a). Beyond horizon graphs: Space efficient time series visualization with composite visual mapping. *IHM 2018 - Actes de la 30ieme Conference Francophone sur l"Interaction Homme-Machine*, 73–82. doi: 10.1145/3286689.3286694

Jabbari, A., Blanch, R., & Dupuy-Chessa, S. (2018b). Composite visual mapping for time series visualization. In *2018 ieee pacific visualization symposium (pacificvis)* (pp. 116–124).

Javed, W., & Elmqvist, N. (2010). Stack zooming for multi-focus interaction in time-series data visualization. *IEEE Pacific Visualization Symposium 2010, PacificVis 2010 - Proceedings*, 33–40. doi: 10.1109/PACIFICVIS.2010.5429613

Javed, W., McDonnel, B., & Elmqvist, N. (2010). Graphical perception of multiple time series. *IEEE Transactions on Visualization and Computer Graphics*, *16*(6), 927–934. doi: 10.1109/TVCG.2010.162

Joyent Inc. (n.d.). *Node.js.* Retrieved 2019-12-18, from `https://nodejs.org`

Julesz, B., & Bergen, J. R. (1983). Human factors and behavioral science: Textons, the fundamental elements in preattentive vision and perception of textures. *Bell System Technical Journal*, *62*(6), 1619–1645.

KAWAI, K., Uchikawa, K., & Ujike, H. (1995). Influence of color category on visual-search. In *Investigative ophthalmology & visual science* (Vol. 36, pp. S654–S654).

Kay, M., Kola, T., Hullman, J. R., & Munson, S. A. (2016). When (ish) is My Bus? User-centered Visualizations of Uncertainty in Everyday, Mobile Predictive Systems. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, 5092–5103. Retrieved from http://www.mjskay.com/papers/chi{_}2016{_}uncertain{_}bus.pdf doi: 10.1145/2858036.2858558

Kim, J., Thomas, P., Sankaranarayana, R., Gedeon, T., & Yoon, H. J. (2016). Pagination versus scrolling in mobile web search. *International Conference on Information and Knowledge Management, Proceedings*, *24-28-Octo*, 751–760. doi: 10.1145/2983323.2983720

Lam, H., Munzner, T., & Kincaid, R. (2007). Overview use in multiple visual information resolution interfaces. *IEEE Transactions on Visualization and Computer Graphics*, *13*(6), 1278–1285. doi: 10.1109/TVCG.2007.70583

Lee, B., Brehmer, M., Isenberg, P., Choe, E. K., Langner, R., & Dachselt, R. (2018). Data Visualization on Mobile Devices. In *Ea '18: Extended abstract proceedings of the acm conference on human factors in computing systems (chi).* ACM.

Limited, A. V. (n.d.). *AirVisual Air Quality Forecast on the App Store.* Retrieved 2020-01-25, from https://apps.apple.com/us/app/airvisual-air-quality-forecast/id1048912974

López, I. F. V., Snodgrass, R. T., & Moon, B. (2005). Spatiotemporal aggregate computation: A survey. *IEEE Transactions on Knowledge and Data Engineering*, *17*(2), 271–286. doi: 10.1109/TKDE.2005.34

Microsoft. (n.d.). *Microsoft Azure - Cloud Computing Services.* Retrieved 2020-01-27, from https://azure.microsoft.com/en-us/https://azure.microsoft.com/pt-pt/

Munzner, T. (2014). *Visualization Analysis and Design*. doi: 10.1201/b17511

Nagy, A. L., & Sanchez, R. R. (1990). Critical color differences determined with a visual search task. *JOSA A*, *7*(7), 1209–1217.

New York Times. (n.d.). *The Upshot.* Retrieved 2019-11-01, from https://www.nytimes.com/section/upshot

Nicolalde, F. D. (2018). Displaying NHP Health Data in Mobile Devices. In *Ea '18: Extended abstract proceedings of the acm conference on human factors in computing systems (chi).* ACM.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606.

Ondov, B., Jardine, N., Elmqvist, N., & Franconeri, S. (2019). Face to face: Evaluating visual comparison. *IEEE Transactions on Visualization and Computer Graphics*, *25*(1), 861–871. doi: 10.1109/TVCG.2018.2864884

Ongwere, T., Connelly, K., & Stolterman, E. (2018). Using ICDMI Model to Guide the Design of Mobile Tool to Support the Care and Treatment of Type-2 Diabetes and Discordant Chronic Conditions. In *Ea '18: Extended abstract proceedings of the acm conference on human factors in computing systems (chi).* ACM.

Paas, F. G. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of educational psychology*, *84*(4), 429.

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27. Retrieved from `https://doi.org/10.1016/j.jbef.2017.12.004` doi: 10.1016/j.jbef.2017.12.004

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision making*, *5*(5), 411–419.

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163. Retrieved from `http://dx.doi.org/10.1016/j.jesp.2017.01.006` doi: 10.1016/j.jesp.2017.01.006

Perin, C., Vernier, F., & Fekete, J. D. (2013). Interactive horizon graphs: Improving the compact visualization of multiple time series. *Conference on Human Factors in Computing Systems - Proceedings*, 3217–3226. doi: 10.1145/2470654.2466441

Pizza, S., Brown, B., McMillan, D., & Lampinen, A. (2016). Smartwatch in vivo. *Conference on Human Factors in Computing Systems - Proceedings*, 5456–5469. doi: 10.1145/2858036.2858522

Plaisant, C., Carr, D., & Shneiderman, B. (1995). Image-Browser Taxonomy and Guidelines for Designers. *IEEE Software*, *12*(2), 21–32. doi: 10.1109/52.368260

Reijner, H., et al. (2008). The development of the horizon graph.

Rideout, V., & Katz, V. S. (2016). Opportunity for all? technology and learning in lower-income families. In *Joan ganz cooney center at sesame workshop.*

Robertson, G., Fernandez, R., Fisher, D., Lee, B., & Stasko, J. (2008). Effectiveness of animation in trend visualization. *IEEE Transactions on Visualization and Computer Graphics*, *14*(6), 1325–1332. doi: 10.1109/TVCG.2008.125

Robinhood Markets, I. (n.d.). *Robinhood: Invest. Save. Earn.* Retrieved 2020-03-19, from `https://apps.apple.com/us/app/robinhood-invest-save-earn/id938003185`

Rosling, H. (2006). *Debunking myths about the "third world"*. Retrieved from `https://www.gapminder.org/videos/ted-talks/hans-rosling-ted-2006-debunking-myths-about-the-third-world/`

Sagi, D., & Julesz, B. (1984). Detection versus discrimination of visual orientation. *Perception*, *13*(5), 619–628.

Sagi, D., & Julesz, B. (1985). " where" and" what" in vision. *Science*, *228*(4704), 1217–1219.

Saito, T., Miyamura, H. N., Yamamoto, M., Saito, H., Hoshiya, Y., & Kaseda, T. (2005). Two-tone pseudo coloring: Compact visualization for one-dimensional data. *Proceedings - IEEE Symposium on Information Visualization, INFO VIS*(November 2005), 173–180. doi: 10.1109/INFVIS.2005.1532144

Sarkar, M., & Brown, M. H. (1992). Graphical fisheye views of graphs. *Conference on Human Factors in Computing Systems - Proceedings*, 83–92. doi: 10.1145/142750.142763

Schneiderman, B. (1996). The Eyes Have It : A Task by Data Type Taxonomy for Information Visualizations Ben Shneiderman University of Maryland. *IEEE Symposium on Visual Languages*, 336–343.

Schwab, M., Hao, S., Vitek, O., Tompkin, J., Huang, J., & Borkin, M. A. (2019). Evaluating pan and zoom timelines and sliders. In *Proceedings of the 2019 chi conference on human factors in computing systems* (pp. 1–12).

Siek, K. A., Rogers, Y., & Connelly, K. H. (2005). Fat finger worries: How older and younger users physically interact with PDAs. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *3585 LNCS*, 267–280. doi: 10.1007/11555261_24

Silver, D. (1994). Feature Visualization. In H. Nielson, Gregory M., Hagen, Hans, Müller (Ed.), *Scientific visualization, overviews, methodologies, and techniques* . (pp. 279–293). IEEE Computer Society, Washington, DC, USA.

Simkin, D., & Hastie, R. (1987). An information-processing analysis of graph perception. *Journal of the American Statistical Association*, *82*(398), 454–465. doi: 10.1080/01621459.1987.10478448

Suri, S., & Watts, D. J. (2011). Cooperation and contagion in web-based, networked public goods experiments. *ACM SIGecom Exchanges*, *10*(2), 3–8.

Treisman, A. M. (1985). Preattentive processing in vision. *Computer vision, graphics, and image processing*, *31*(2), 156–177.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive psychology*, *12*(1), 97–136.

Treisman, A. M., & Gormican, S. (1988). Feature analysis in early vision: evidence from search asymmetries. *Psychological review*, *95*(1), 15.

Tufte, E. R. (1990). *Envisioning information*. doi: 10.2307/3325378

Tufte, E. R. (2001). *The Visual Display of Quantitative Information* (Vol. 2 ed.). Cheshire, CT: Graphics press.

Tukey, J. W. (1977). *Exploratory data analysis* (Vol. 2). Reading, MA.

Tuovinen, J. E., & Paas, F. (2004). Exploring multidimensional approaches to the efficiency of instructional conditions. *Instructional science*, *32*(1-2), 133–152.

Walker, J., Borgo, R., & Jones, M. W. (2016). TimeNotes: A Study on Effective Chart Visualization and Interaction Techniques for Time-Series Data. *IEEE Transactions on Visualization and Computer Graphics*, *22*(1), 549–558. doi: 10.1109/TVCG.2015.2467751

Ware, C. (2019). *Information visualization: perception for design*. Morgan Kaufmann.

Watson, B., & Setlur, V. (2015). Emerging research in mobile visualization. In *Proceedings of the 17th international conference on human-computer interaction with mobile devices and services adjunct* (pp. 883–887).

Weigle, C., Emigh, W. G., Liu, G., Taylor, R. M., Enns, J. T., & Healey, C. G. (2000). Oriented texture slivers: A technique for local value estimation of multiple scalar fields. In *Proceedings graphics interface* (pp. 163–170).

Wickens, C. D., Hollands, J. G., Banbury, S., & Parasuraman, R. (2015). *Engineering psychology and human performance*. Psychology Press.