Designing and Evaluating Feedback Schemes for Improving Data Visualization Performance

A Major Qualifying Project Report

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

WPI

In partial fulfillment of the requirements for the

Degree of Bachelor of Science

By

Meixintong Zha

Submitted to

Professor Lane T. Harrison

Worcester Polytechnic Institute

Abstract

Over the past decades, data visualization assessment has proven many hypotheses while changing its platform from lab experiment to online crowdsourced studies. Yet, few if any of these studies include visualization feedback participants' performance which is a missed opportunity to measure the effects of feedback in data visualization assessment. We gathered feedback mechanics from educational platforms, video games, and fitness applications and summarized some design principles for feedback: inviting, repeatability, coherence, and data driven. We replicated one of Cleveland and McGill's graph perception studies where participants were asked to find the percentage of the smaller area compared to the larger area. We built a website that provided two versions of possible summary pages - with feedback (experimental group) or no feedback (control group). We assigned participants to either the feedback version or the no feedback version based on their session ID. There were a maximum of 20 sets of twenty questions. Participants needed to complete a minimum of 2 sets, and then could decide to either quit the study or continue practicing data visualization questions. Our results from a 64 participants study suggest that, on average, the feedback group may have improved slightly faster than the no feedback group.

Content

Chapter 1 Introduction

Data visualization is the graphical representation of data which involves producing images that communicate relationships among the represented data to viewers. People can distinguish differences in line length, shape, orientation, and color (hue) readily without significant processing effort; these are referred to as "pre-attentive attributes". For example, it may require significant time and effort ("attentive processing") to identify the number of times the digit "5" appears in a series of numbers; but if that digit is different in size, orientation, or color, instances of the digit can be noted quickly through pre-attentive processing [6].

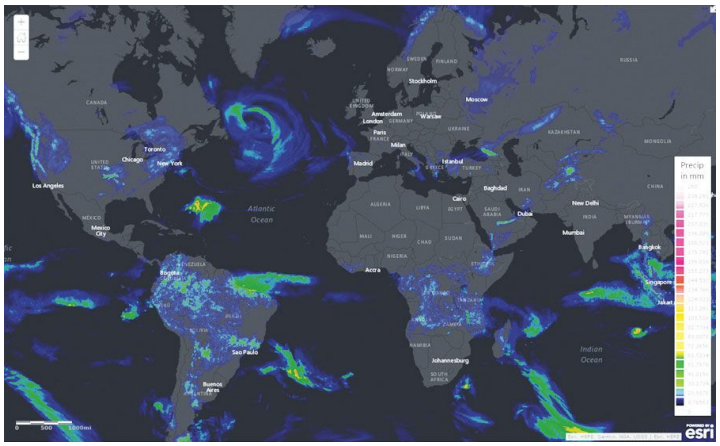

Figure 1a: Predicting Weather with advanced CSI at

https://www.esri.com/about/newsroom/arcnews/predicting-the-weather-with-advanced-gis/

In today's world, data visualization is applicable to different aspects of life. Companies rely on data visualization to outline the correlations of a large data set. Meteorologists apply dynamic mental model data visualization to produce weather forecasts with higher accuracy [17] (Example in Figure 1a). Data visualization also holds a critical role in healthcare, because it

takes large data sets and turns them into an easily consumable format for customers to overview [13]. Through publicly available data, data visualization can illustrate public transportation data to help us make optimal travel plans(Figure 1b). For instance, WPI graduates Michael Barry and Brian Card produced the following visualizations using data captured from the MBTA (Massachusetts Bay Transportation Authority) public data for the entire month of February, 2014. These graphs help us see how the MBTA system operates on a daily basis, how people use the system, how that affects the trains and also how this ties back to people's daily commute. (http://mbtaviz.github.io/).

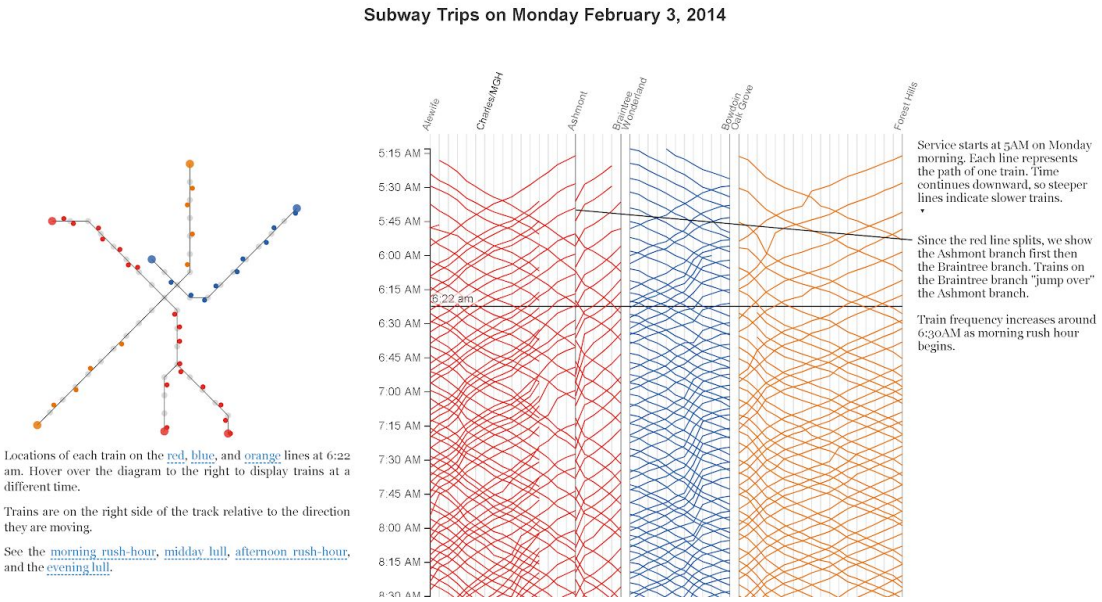

Figure 1b: MBTA Train Map(Left); A Marey's Train graph for MBTA(Right).

For decades, data visualization scientists have conducted numerous experiments to quantify and model people's performance on visualization literacy.. To name a few, the Cleveland and McGill's data visualization perception experiment tested people's accuracy of stating the percentage of the smaller labeled division compared to the larger labeled division in a

graph [4]. Cleveland and McGill had found a hierarchy for the data visualization properties which will be listed in the background chapter. Heer and Bostock conducted a replication of Cleveland and McGill's graphic experiment on Amazon mechanical turk to ensure the credibility of crowdsourcing for data visualization experiments [10]. Harrison *et al.* used JND (Just Noticeable Difference) method to produce a perceptually-driven ranking chart for correlation data [9]. These experiments tested people's perception and cognitive functioning in data visualization literacy without influencers such as educational tips, answer correction, or feedback.

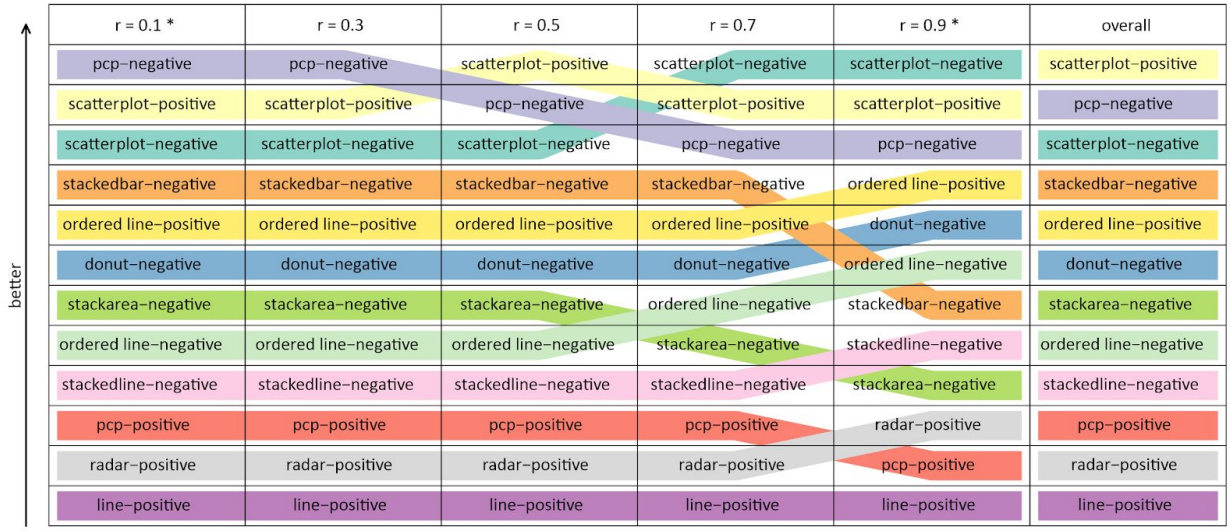

Figure 1c: Harrison *et al.* produces a perceptually-driven ranking for individual correlation (r) values, as well as an overall ranking (right column).

Recently, more and more data visualization literacy studies are hosted online for a number of possible reasons. Online crowdsourcing platforms can provide up to an order of magnitude cost reduction; such savings can be invested into more subjects or more conditions [10]. Researchers can gain access to a wider population across the world with crowdsourcing. The side effect is that online experiments, unlike traditional experiments, participants were not

given face-to-face consent, brief, or debrief instructions. The participants were also not monitored, which means they could take intermission from the studies. This leads us to consider the mechanics for the design of the online experiment user interface so that participants can have a more immersed experience.

Imagine you are requested by your friend to finish an online data visualization experiment, you went up to the website and hope to finish this task as soon as possible. You signed the consent form and started the experiment. There were more questions than you expected, and the questions were actually difficult. Even though you balanced your pace with a good level of confidence in your correctness rate, the boredom still kicked in. You finally finished the experiment. However, if there was no feedback on your performance, how would you feel about the time and effort you have spent on the experiment?

Depending on the number of the question, some people don't have the motivation to carefully go through all the questions. This will result in inaccuracy in the analysis of people's optimal graph reading ability. In Harrison's affective priming experiment, people with positive emotions have better graphical reasoning performance as well as higher improvement rate than people with negative emotions [8]. A data visualization experiment design should incorporate a Mood-Induction Procedure to enhance participants' positive emotions, so that they have a better chance with improvement on data visualization reading in the future. Mood-Induction Procedure usually involves providing visual or auditory stimulation to influence a subject's mood.

A problem that we neglect in the past data visualization is using feedback as a tool to optimize users' performance. Online experimentation has the capability to generate logical and scripted feedback. In the educational field, quantitative studies show that feedback assessment

techniques promote learning and correct mistakes. Feedback gives students areas to improve as well as confidence. Switching back to data visualization science, there weren't many studies focusing on the effect of feedback on user performance.

To understand feedback's potential effects, we explored the effectiveness of visual feedback techniques as well in improving subjects' data visualization reading performance by reviewing related studies and gathering feedback samples. We tested our hypothesis by conducting an online experiment which assesses user data visualization comprehension performance (recreated from Cleveland and McGill) by building a website that provides two versions of break page - feedback versus no feedback. This experiment was an online study that contained 20 sets of twenty questions. The twenty questions consisted of five questions from each of the bar charts, pie charts, bubble charts, and stacked bar charts. Users had to complete at least 2 sets of questions before quitting, so we can observe their performance improvements. Results of a 64 participants study show that, on average, the feedback group improved slightly faster than the no feedback group. Also, each chart type had a different improvement rate.

The rest of this report is organized as below. We looked at data visualization related studies, the MOOC (Massive Open Online Course) studies, and psychological theories to search for potential feedback mechanics and performance improvement techniques in Chapter 2. Some feedback mechanics were gathered from educational platforms, video games, and fitness applications to learn the feedback design principles in Chapter 3. We analyzed the data by the whole group and two special cases in Chapter 4. The results and future implementations were discussed in Chapter 5.

Chapter 2 Background

Quantitative user performance studies on data visualization have been conducted. Later data visualization researchers adopted Cleveland and McGill's divided region comparison experiment to test their hypothesis. This chapter explores the related studies about data visualization, MOOC designs, and psychological conditions to find potential feedback models to improve user performance in online data visualization studies.

2.1 Related Studies for User Performance in Data Visualization

2.1.1 Cleveland and McGill's Graphical Perception Experiments

In 1984, Cleveland and McGill published their findings on the general hierarchy of data visualization properties which users most accurately understand. They applied two approaches. The first approach Cleveland and McGill took is to identify a set of elementary perceptual tasks that are carried out when people extract quantitative information from graphs. The second approach is to order the tasks on the basis of how accurately people perform them.

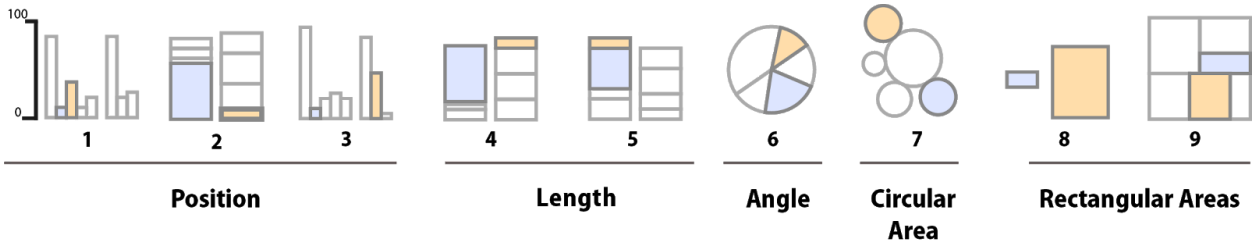

Figure 2.1.1

Cleveland and McGill found that participants had more accurate reading in bar graphs than pie charts consistently. Figure 2.1.1 shows the hierarchy list of data visualization property lists [4]:

1. Position along a common scale (bar chart, dot plots)
2. Positions along nonaligned, identical scales (small multiples)
3. Length, direction, angle (pie chart)
4. Area (treemap)
5. Volume, curvature (3-D bar charts, area charts)
6. Shading, color saturation (heat maps, choropleth maps)

For this project, we will replicate Cleveland and McGill's comparing the two areas in different chart type experiments. In Cleveland and McGill's experiment, participants were tested on different graph types without receiving feedback. This led us wondering how feedback would affect the original results. For example, would feedback help users improve more on one certain graph type over another?

2.1.2 Heer and Bostock's Crowdsourcing Graphical Perception Experiments

Heer investigated whether online crowdsourced platforms could be adequate for graphical perception research. He assessed the feasibility of using Amazon's Mechanical Turk to

evaluate visualizations. He replicated prior laboratory studies on spatial data. Encodings and luminance contrast using crowdsourcing techniques [10]. The matching results suggest that crowdsourcing is viable for testing graphical perception. Another finding from Heer was that the qualification tasks and verifiable questions help ensure high-quality responses and that experimenters can accelerate the time to results by increasing the compensation level [10]. Heer's study helps us to verify the quality of responses from crowdsource websites. Our project will use Prolific.co to recruit participants.

3.1.3 Harrison's Visualizations ranked by Weber's Law

Harrison conducted a large-scale (n=1687) crowdsourced experiment on using Weber's Law as a tool to rank the perception of correlation in nine commonly used visualizations. Figure XX shows Weber's Law stated that the size of the difference threshold appeared to be lawfully related to initial stimulus magnitude[20].

$$\frac{\Delta I}{I} = k$$

where ΔI (delta I) represents the difference threshold, I represents the initial stimulus intensity and k signifies that the proportion on the left side of the equation remains constant despite variations in the I term

Figure 3.1.3

He found that for all tested visualizations, the precision of correlation judgment could be modeled by Weber's law. However, correlation judgment precision showed striking variation between negatively and positively correlated data in parallel coordinate plane graphs. This suggested that these symmetries might be related to the visual features participants attend to when judging correlation [9].

2.2 Psychological Factors

How well the participants perform in data visualization tasks depends on numerous factors which include academic background, graph literacy experience, good focus, and etc. This section will talk about performance factors related to psychology.

2.2.1 Participant Motivation

Maslow's hierarchy of needs is a theory in psychology proposed by Abraham Maslow in his 1943 paper "A Theory of Human Motivation" in Psychological Review [14].

Maslow's hierarchy of needs is used to study how people intrinsically partake in behavioral motivation. Figure 3 shows a pyramid of Maslow's hierarchy of needs. This means that in order for motivation to arise at the next stage, each stage must be satisfied within the individual themselves.

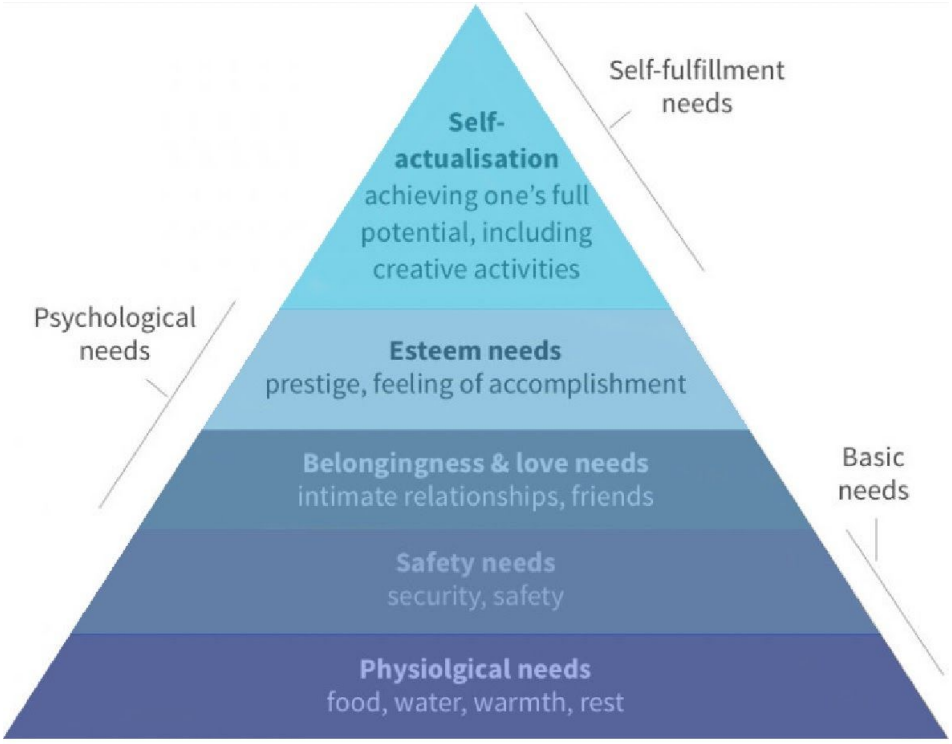

Figure 3: *Maslow's Pyramid of Needs*, created by Chiquo.

We can improve participant motivation on participating data visualization experiments by satisfying their needs according to Maslow's theory. First, we could provide monetary reward for taking the experiment. This is the first level of Maslow's Hierarchy. For example, Heer's online data visualization experiment concludes that with higher compensation level leads to higher quality performance. We could give participants an instruction to take this experiment in a quiet and safe room which satisfies the second level of the pyramid.

Next, we could give participant feedback on their performance rankings among all the other participants to create a sense of belongingness, the third level of Maslow's Hierarchy.

We should emphasize on providing positive feedback so that participants can develop solid self-esteem on data visualization literacy. This is the fourth level of Maslow's Hierarchy. We

could provide immediate feedback to participants so that participants could improve their performance. This is the fifth level of Maslow's Hierarchy.

2.2.2 Attention Span

Attention span is the amount of concentrated time a person can spend on a task without becoming distracted [2]. The distraction occurs when the individual is uncontrollably drawn to some other activity or sensation [19]. There are two types of attention span: transient attention and selective sustained attention. Transient attention is a short-term response to a stimulus that temporarily attracts/distracts attention. Selective sustained attention, the attention for most online experiments, is the consistent attention fixated on a single task. Common estimates of the attention span of healthy teenagers and adults range from 10 to 20 minutes [21]. Attention restoration theory (ART) asserts that people can concentrate better after spending time in nature, or even looking at scenes of nature. Natural environments abound with "soft fascinations" which a person can reflect upon in "effortless attention", such as clouds moving across the sky, leaves rustling in the breeze or water bubbling over rocks in a stream [11].

2.2.3 Memory Span

One cognitive limitation humans have is memory span. Memory span refers to the longest list of items (e.g., digits, letters, words) that a person can repeat back in the correct order on 50% of trials immediately after presentation. Miller observed that the memory span of young adults is approximately seven items,and memory span is not limited in terms of bits but rather in

terms of chunks. A chunk is the largest meaningful unit in the presented material that the person recognizes—thus, what counts as a chunk depends on the knowledge of the person being tested. For instance, a word is a single chunk for a speaker of the language but is many chunks for someone who is totally unfamiliar with the language and sees the word as a collection of phonetic segments [15]. When we design a summarizing feedback page, we should consider the number of items to highlight so that users could retain the important information.

2.3 MOOC

We looked at MOOC studies, because, similar to online crowdsourcing, MOOC uses online servers as platforms to attract students and teachers. MOOC stands for Massive Open Online Course. These MOOCs are based on traditional university courses. The advantage of MOOC is that they significantly broaden the number of students who can be exposed to university-level courses with lower cost and no requirement of commute.The disadvantage of MOOCs is that Critics argue that MOOCs are inferior to the university courses they mimic because they eliminate teacher-student interactions and involve limited student-student interactions. Some famous MOOC platforms are edX, Coursera, and Udacity.

2.3.1 Reinecke's Demographic MOOC Experiment

Reinecke conducted an empirical study of how students navigate through MOOCs based on students' age and country of origin. She performed data analysis on the activities of 140,546 students in four edX MOOCs and found that certificate earners skip on average 22% of the course content, that they frequently employ non-linear navigation by jumping backward to earlier lecture sequences, and that older students and those from countries with lower student-teacher ratios are more comprehensive and non-linear when navigating through the course. These results suggest design recommendations for MOOC platforms to develop more detailed forms of certification that incentivize students to put more effort rather than just doing the minimum necessary to earn a passing grade[18]. If we compare this study to Heer and Bostock's finding (Section 2.1.2), there are similarities over incentives which the incentives tend to influence participants to finish the experiment/study faster. This is important to keep in mind for designing the experiment, because, on a crowdsourcing platform, the participants are paid to complete an experiment.

2.3.2 Clarà and Barberà's Connectivism in MOOCs

Clarà and Barberà reflected on related MOOC and psychology studies; they discussed the connectivist conception of learning in Web 2.0 environments, which underpins the pedagogy of what are known as cMOOCs (connectivist massive open online courses). Connectivist pedagogy indicates that material should be aggregated, remixable, repurposable, and should be targeted at future learning [5]. They thought that connectivism does not provide an adequate explanation of

learning phenomena in MOOC platforms, and therefore it is not able to provide an adequate pedagogy for MOOCs [3]. This study suggested to limit participants' interactions to consolidate a traditional learning experience for an online educational website design.

Chapter 3 Design and Methodology

3.1 Reviewing The Design Space for Delivering Feedback to End-Users

3.1.1 Educational Applications Feedback Mechanics

The emerging usages of personal computers and smartphones evolve education to be more remotable and efficient with the help of different types of feedback. In this section we will look at different types of educational softwares and platforms such as Udemy, Lynda, Class Dojo, Lab In The Wild. We will then conduct an in-depth review on Duolingo - one of the highest rating language learning applications. In the end, we will summarize the commonalities of these educational applications.

3.1.1a Udemy and Lynda - Educational Platforms

Udemy is an American online learning platform aimed at professional adults and students, developed in May 2010. The platform has more than 50 million students and 57,000 instructors teaching courses in over 65 languages as of Jan 2020 (udemy.com). Lynda is an American website offering video courses taught by industry experts in software, creative, and business skills, founded in 1995(lynda.com). Both Udemy and Lynda give participants certificates as incentives. Since online courses have become popular, acquiring skills with visible approval are more attractive to users than self learning without feedback. The certificates contain

information such as course title, date, username, and time spent on the course which give users both satisfaction and virtual copies of achievement.

3.1.1b Class Dojo - Educational Communication System

Class Dojo is an online behavior management system intended to foster positive student behaviors and classroom culture. Class Dojo system utilizes 'Dojo Points' to motivate students to practice good classroom behavior. Students and teachers can also post videos and photos to show off their class projects and highlight moments. In a digital age where kids have their own cell phones at a very young age, Class Dojo provides an interactive social media for the students.

3.1.1c Lab In The Wild - Online Crowedsource for Research Studies

Lab In TheWild tests participants' abilities and preferences. At the end of each experiment, participants will see a page with their personalized feedback, which let them compare themselves to other people around the world. The designers of Lab In The Wild chose to use a competitive ranking system to elevate participants' motivation to get better feedback thus leading to more accurate participant input. This system might promote higher likelihood of participants answering experimental questions with their full abilities.

3.1.2 Duolingo Case Study

Duolingo is a language learning platform that has both a website and phone app version with over 300 million users (duolingo.com). Duolingo provides 30 languages for users to choose from. It is voted the best educational app since 2013[Gigaoam]. Attracted by its popularity and praises, we would like to take a deep review on all Duolingo's features and feedback mechanisms.

3.1.2a Duolingo Features

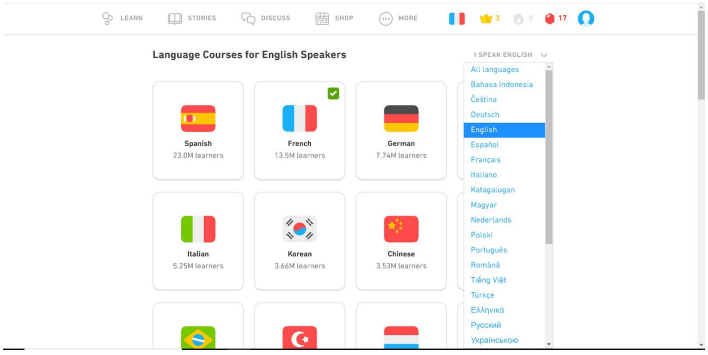

Figure 3.1.2a: Duolingo's language selection

After registration, users can choose which language they want to learn as well as set the system language. Later, users are able to change the language any time they want by clicking the nationality flag on the navigation bar.

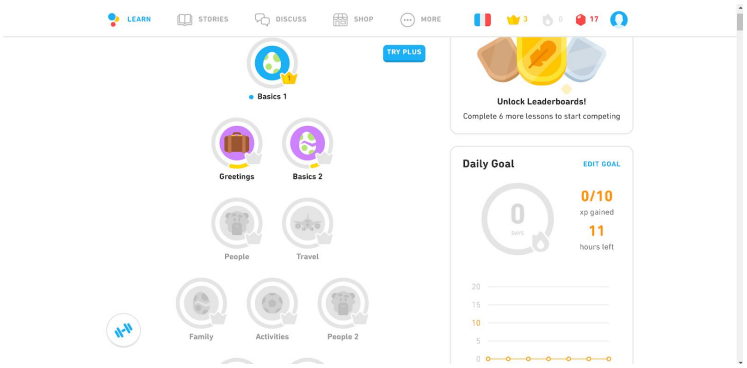

Figure 3.1.2b: Duolingo's main interface

In the learning page, the courses are separated into small modules arranged from easiest on top and hardest in the bottom. Users have to finish all modules in one row in order to move to the next row. Each module contains five levels. Each module takes about 5 - 20 lessons to complete. Once users complete a lesson, the donut chart of the corresponding module will show the progress.

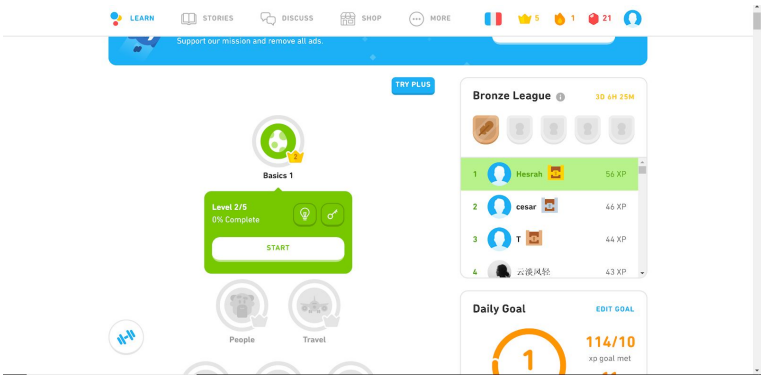

Figure 3.1.2c: Duolingo's ranking system

Duolingo uses a ranking system which assimilates to sports and video games. The Duolingo ranks are: Bronze, Silver, Gold, Sapphire, Ruby, Emerald, Amethyst, and Pearl. The promotion and demotion occurs on a weekly basis. By the end of all competitions, those who

keep their ranks within the promotion zone are promoted one rank higher. Shown is Figure 3.1.2c, on the right, you can see there is a box labeled Bronze League which shows your placements among all the participants in the Bronze League.

Figure 3.1.2d: Duolingo's immediate feedback samples

Duolingo uses immediate feedback after every single question. If the user makes a mistake, Duolingo will provide the correct answer. Duolingo also tells you how many questions you did correctly in a row. If you get a question wrong. The same question will appear in this lesson again. The user has to answer every question correctly once in order to pass this section.

3.1.2b Duolingo's Feedback Mechanics Compare to Video Games'

Duolingo applies several design and feedback mechanics which make language learning entertaining thus attracting more users. Firstly, Duolingo gives users freedom on choosing target language and daily study goals. Secondly, Duolingo separates traditional one hour language courses into multiple small sections of 10 to 20 questions. Thirdly, Duolingo applied immediate feedback (visual and audio) so that users can learn from their mistakes. Lastly, Duolingo applied

cute animated characters to make learning fresh and fun. We asked our computer science major friend to test out Duolingo. She was a Nintendo DS and Playstation Portable gamer. She loved playing RPGs (Role-playing games) such as Pokemon, Final Fantasy and farm simulation games such as Harvest Moon. For an hour of studying French on Duolingo's website, she said she felt like she was playing a RPG game instead of taking an online course. Each lesson took only about three minutes to run through. The lessons were not difficult. She got experience points from each lesson that she completed. This reward system Duolingo applies tends to occur in many video games which makes users addicted to grinding; in this case, promotes longer language learning time.

3.1.3 Video Game Feedback Mechanics

Video games are one of the most popular hobbies nowaday. There is even a psychological disorder called Internet Gaming Disorder in DSM5 (Diagnostic and Statistical Manual of Mental Disorders)[dsm]. Pro game players can spend more than three continuous hours on gaming. What makes video games so popular and addicting? In this section, we would explore feedback mechanics in video games.

Feedback is the cornerstone for players to learn game mechanics and control. Even with a large variety of game genres, most video games tend to use both immediate feedback and summarizing feedback to consolidate gaming experience. We investigated some popular genres and listed their corresponding feedback:

Game Genre	**Immediate Feedback**	**Summarizing Feedback**

FPS(First Person Shooter)	screen shake effect, bleeding effect, ammo counts, sounds effect...	end game page, music
RTS(Real Time Strategy)	dialogues, map visibilities, time counter	end game page, music
MOBA(Multiplayer Online Battle Arena)	character sound, damage number on enemy's head, map visibility, hp bars	user rank points, end game page,
RPG(Role Playing Game)	hp bars, bleeding effect	success/defeat music, end fight page
MMO(Massively Multiplayer Online)	screen effects, sound effects	a summary page, music
MMORPG(Massively Multiplayer Online Role Playing Games)	character level, sound effect	user rank points, success/defeat music
Puzzle	hints, sound effect	level completion page with stat

We found that video games tend to use immediate feedback both on visual and audio to add immersion, improve user experience, and help users learn game mechanics. Most game genres provide users a summarizing feedback page which either presents users with their performance statistics or simply provides users incentives such as gold and experience points. We looked at feedback pages from League of Legend, Overwatch, HearthStone, Fortnite, Fifa, Flow. These games have an overall dark color scheme results pages. They used radar charts, time charts, a game map that has dots which shows the number of kills by players, tables listing performance such as kills and gold. A lot of competitive online multiplayer games have third parties making detailed summarizing feedback pages for users to improve their performance. Websites such as OP.GG and Mobalytics are examples of such third party websites.

Figure 3.1.3a: OP.GG screenshot. It shows all players's ranking, performance, KDA(kill, die, assist), damage, and item choices.

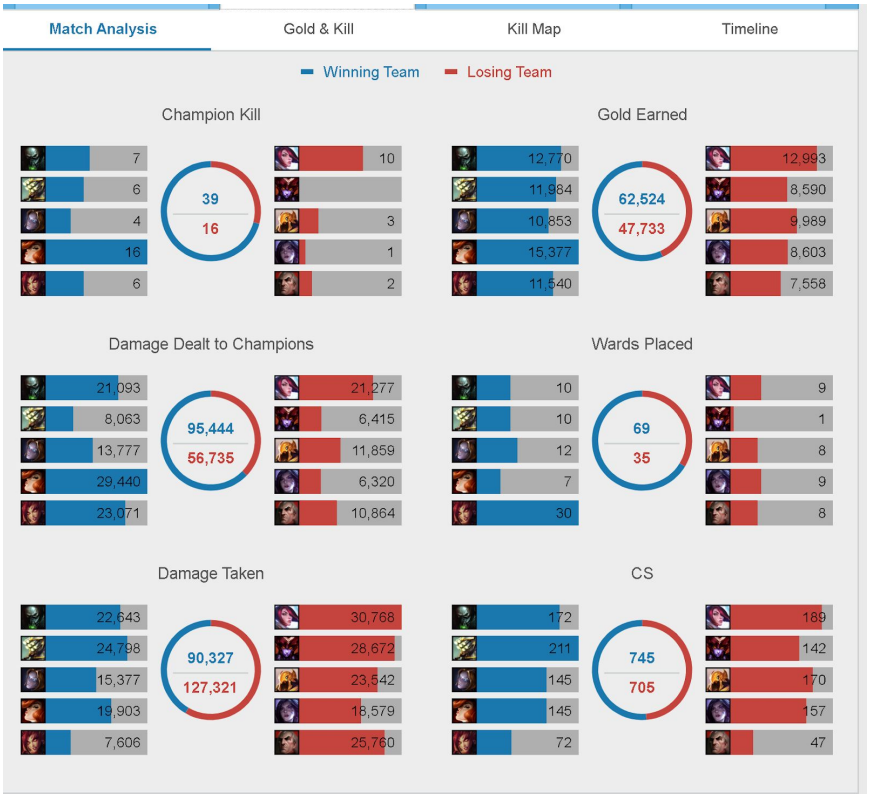

Figure 3.1.3b: OP.GG uses donut charts analysis for champion kill, gold earned, damage death to champions, ward placed, damage taken, and cs.

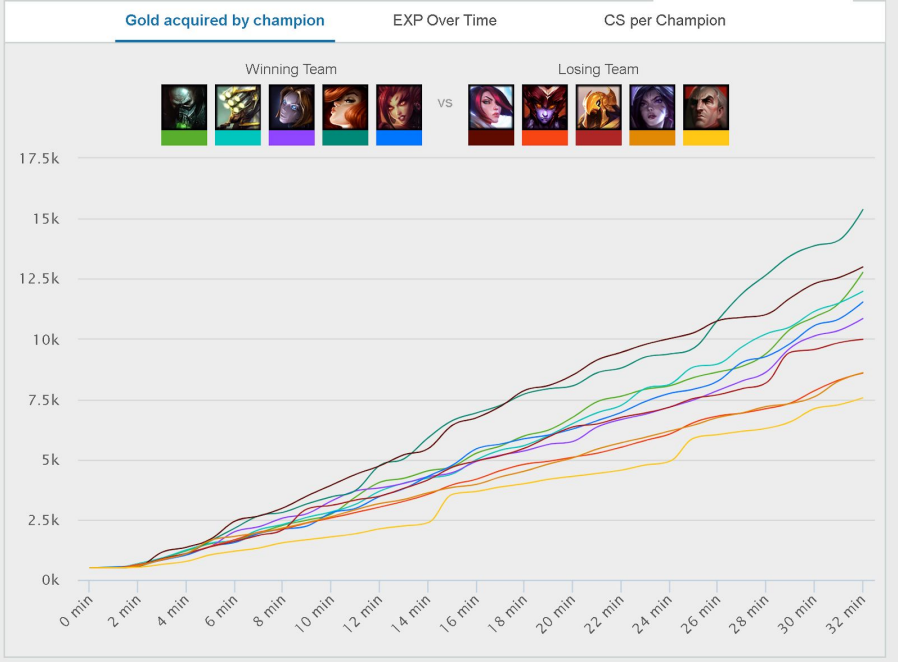

Figure 3.1.3c: a bar chart of gold acquired over time for each player.

3.1.4 Fitness Application Feedback Mechanics

With the emerging market for health and fitness technology, as of 2015, there were about 165,000 fitness applications in the market [16]. We investigated popular fitness applications such as Fitbit, Google fit, My Fitness Pal, Lumosity (a brain fitness app), and Pedometer. We found recurring usages of data visualization like donut charts and bar charts to show statistics such as active level, calories burnt, and duration of workout. Fitness apps also utilize refreshing color schemes to depict a healthy mood. Most fitness apps such as Google fit and Pedometer let users

set their personal goals just like Duolingo. This gives users a sense of freedom and encourages users to track their meals and exercises.

3.1.5 Design Principles

Through conducting research online [1][12] and reviewing some of the successful feedback examples from educational websites/applications, video games, and fitness applications, we listed some of the important design principles for a feedback page:

- **Inviting**: The feedback should be an affirmative influencer which the user wants to improve his/her performance in order to get positive feedback.
- **Repeatability**: There should be a repeating occurrence of feedback where the user consistently receives the feedback after a certain amount of exercise.
- **Coherence**: The content of the feedback must relate to the context. Coherence is imperative to avoid confusion while assisting users to improve their performance.
- **Data Driven**: The feedback has to be objective and real.

3.2 Replicating and Expanding Visualization Experiments to Include Feedback

For this project, we replicated Cleveland and McGill's comparing two areas of different chart types. Participants were given questions to answer them to predict what is the percentage of the smaller area compared to the bigger area (see Figure 3.2.1a).

3.2.1 Question Generation

We generated 20 sets of questions. Each set contains questions about barchart, bubble charts, pie charts, and stacked bar charts as shown in Figure 3.2.1a. Participants were given five questions for each graph type in one set. Shown in Figure 3.2.1b, the five questions each corresponded to 10%, 25%. 50%, 75%, and 90% comparison between the two areas.

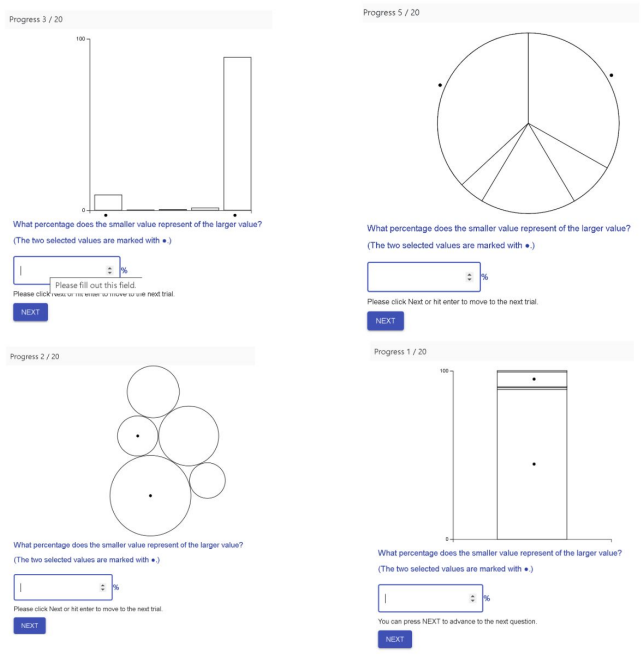

Figure 3.2.1a: examples of data visualization questions from this project's testing website.

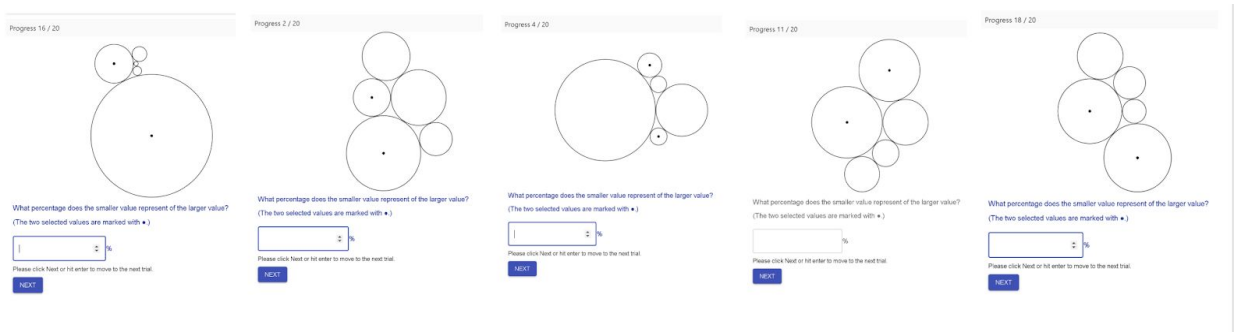

Figure 3.2.1b: example of the 10%, 25%, 50%, 75%, and 90% area ratio of bubble charts within one set of 20 questions.

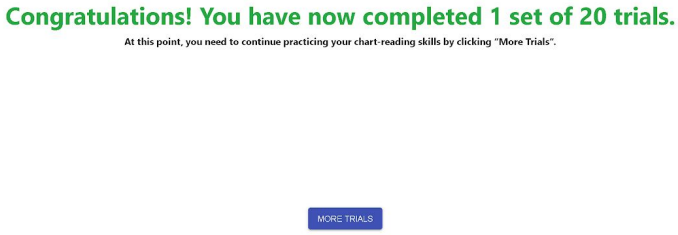

Figure 3.2.1c: the no feedback summary page of the testing website.

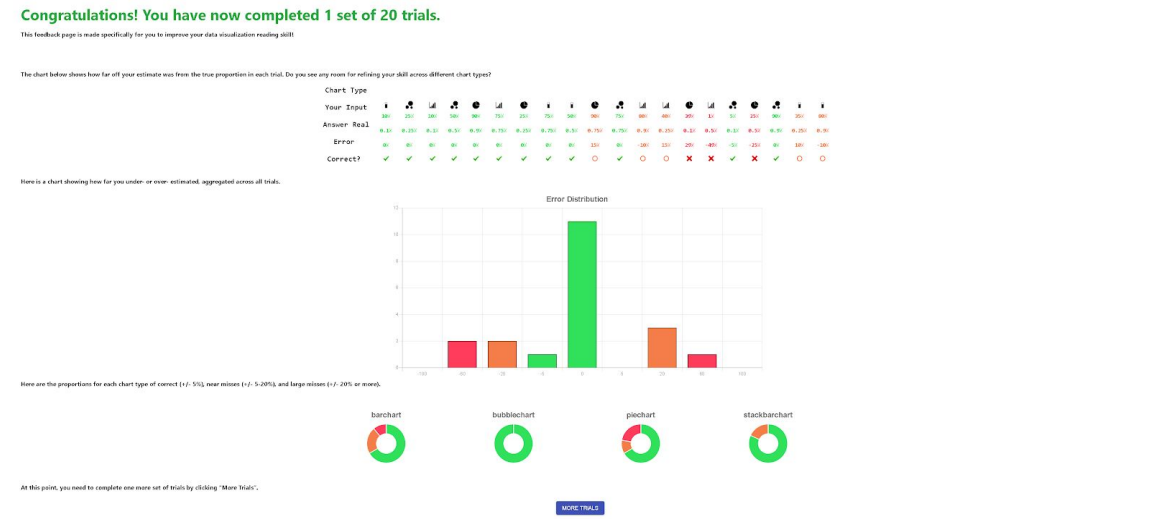

Figure 3.2.1.d: the feedback summary page of the testing website. Participants received three types of feedback: 1) performance on each trial, 2) graph of error of distribution, and 3) donut charts for each chart type.

3.2.2 Feedback Design

We provided three types of feedback that were corehent and data driven.

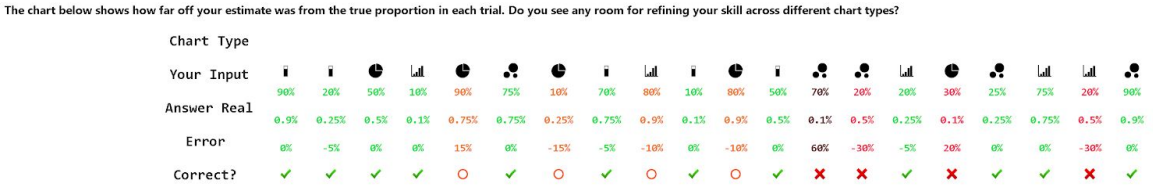

Figure 3.2.2a

The first feedback was a trial by trial performance feedback. Figure 3.2.2a shows that the feedback was in table format which provided chart type, participant's input, real answer, error, and whether the participant answered the problem correctly. A color scheme of green, orange, red, and brown is applied for making the incorrect questions conspicuous to the participants so they could pay attention to the corresponding chart type questions in the next set. Correct (green) ranged from 0% to ±5%; near misses (orange) ranged from positive or negative 5% to 20%; large misses (red) ranged from positive or negative 20% to 60%; >±60% (brown) was counted as extreme misses.

Figure 3.2.2b

The second feedback was a bar graph of error distribution. The color scheme corresponded to the first feedback table. Participants could read this graph and understand the offsets of their answer to the correct answer. Figure 3.2.2b shows a participant who tended to make smaller estimates compared to the correct answer.

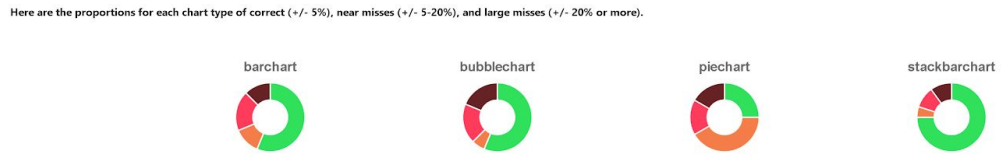

Figure 3.2.2c

The third feedback was a bar graph of proportion of each chart type of correct, near misses, large misses, and extreme misses. The color scheme corresponded to the previous two feedbacks. We chose donut charts because donuts charts were frequently implemented in video games and fitness apps.

3.2.3 Flow of Experiment

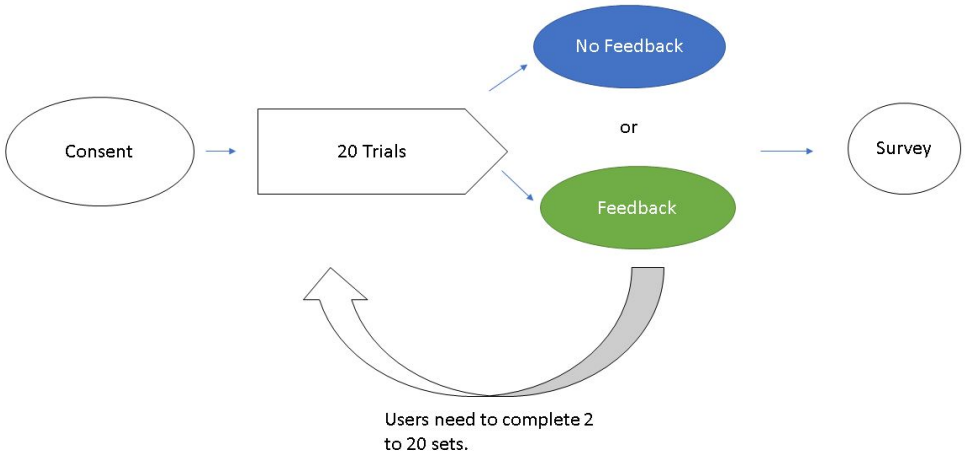

Figure 3.2.3a: flow of experiment diagram

The study utilized Prolific (Prolific.co) - a crowdsourced research website - to recruit participants. After participants agreed to the consent form (see Appendix A), they could start the study. The study contains 400 questions which are divided into 20 twenty-question-set. Each question takes about five to twenty seconds to finish. A test set contains twenty questions, so it takes about two to five minutes to finish one question set. The participants were required to complete at least two sets of questions before ending the experiment. The whole test should take 40 minute to 1hr 40 minutes to complete. After participants finished their desired number of question sets, they needed to fill out a survey which contains demographic and opinion questions(see Appendix B, C).

3.2.4 Experiment Expectation

We expected to have 60 participants. We would divide the 60 participants into two groups of 30 participants. Each participant was assigned a session ID (from 1 to 60). If the session ID is odd, the participant will take the no feedback website. If the session ID is even, the participant will use the feedback website.

We expect the feedback group to have higher performance improvement than no feedback group. For the first question set, both groups' performance should show their knowledge of data visual literacy based on their past experience. For the following question sets, the feedback group should improve faster than the no feedback group, because the feedback page will tell them about their performance on each graph type which makes them more aware of the accuracy of their answers. No feedback group participants will receive any feedback on their performance; the absence of information and motivation for improvement will result in slower improvements on data visualization literacy.

Chapter 4 Results

4.1 User Demographic

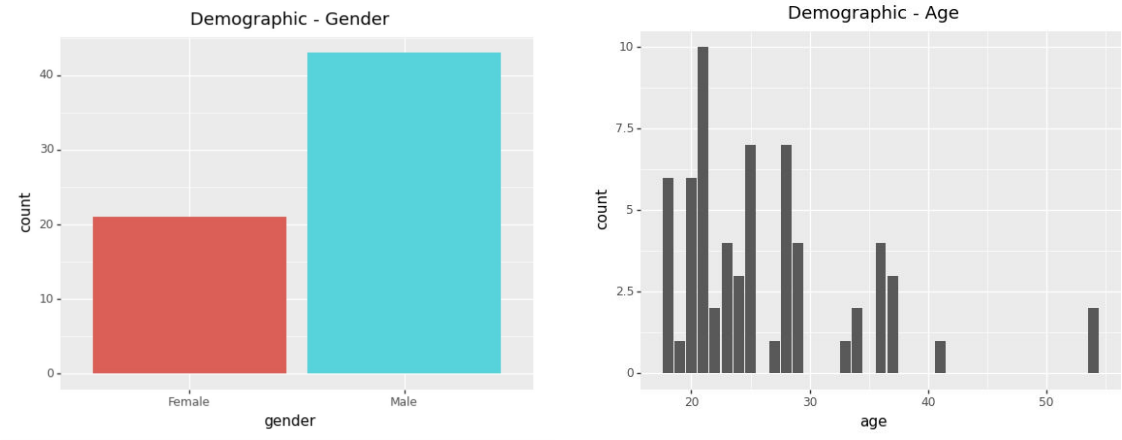

Figure 4.1: Demographic data on gender(on the left); demographic data on age(on the right).

We recruited 64 participants recruited from Prolifics crowdsource website. Figure 4.1 shows that 43 of the participants were male; 21 of the participants were female. The youngest participants were 18 and the oldest participants were 54. The average age was 26 years old.

4.2 Participants Performance Analysis

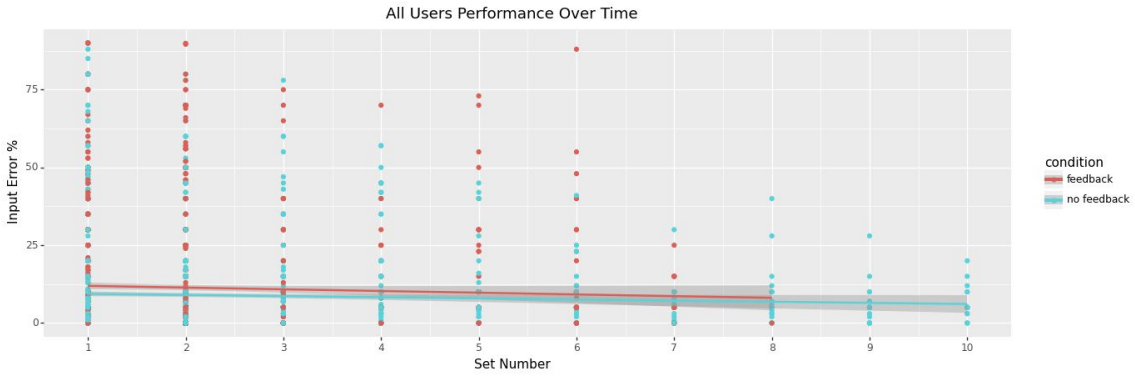

Figure 4.2a: All 64 participants' performance data on scatter points diagram with local regression smooth fit.

Figure 4.2a shows a scatter diagram of participants' judgemental error across the number of sets over all 64 participants without filtering. There are two lines which each represent the local regression smooth for participants with or without feedback; we will treat these lines as participants' learning curves. The red dots represent participants' inputs with feedback; the blue dots represent participants' inputs without feedback. The x-axis represents the set number. For the experiment, there is a maximum of 20 sets of trials. The highest number x ticks label is 10 which means the highest amount of sets participants completed is 10 sets. However, red dots only exist until the 8th set, which means the maximum number of trials completed by participants with feedback was eight. The y-axis represents the offset of participant input error. The input error % is calculated by:

$$input\ error\% = |participants\ \ input - correct\ answer|\ \%$$

This graph shows that the non feedback group completed more sets than the feedback group while there was no definitive difference between two groups' learning curves.

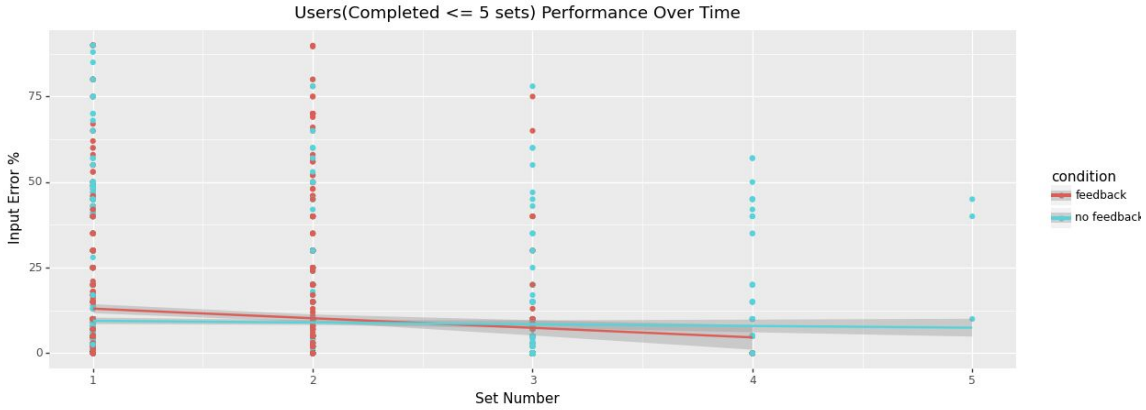

Figure 4.2b: A scatter points diagram for participants completed less than or equal to 5 sets with local regression smooth fit.

We found that there were only three out of the 64 participants who completed more than five sets. In order to focus on the majority of the participants , we looked at data focusing on participants who completed less than or equal to five sets. Figure 4.2b shows a scatter diagram of participants' judgemental error across the number of sets. We can see that the feedback group's average judgemental error was about 12.5% at set one; it reduced to about 5% at set four. The no feedback group average judgemental error was about 10%; it reduced to about 8% at set five.

$$feedback\ group\ improvement\ rate\text{: } \frac{12.5\% - 5\%}{3\ sets} = 2.5\%\ per\ set$$

$$no\ feedback\ group\ improvement\ rate\text{: } \frac{10\% - 8\%}{4\ sets} = .5\%\ per\ set$$

This data shows that the feedback group has 2% per set improvement rate faster than the no feedback group.

4.2.1 Analysis for Each Chart Type

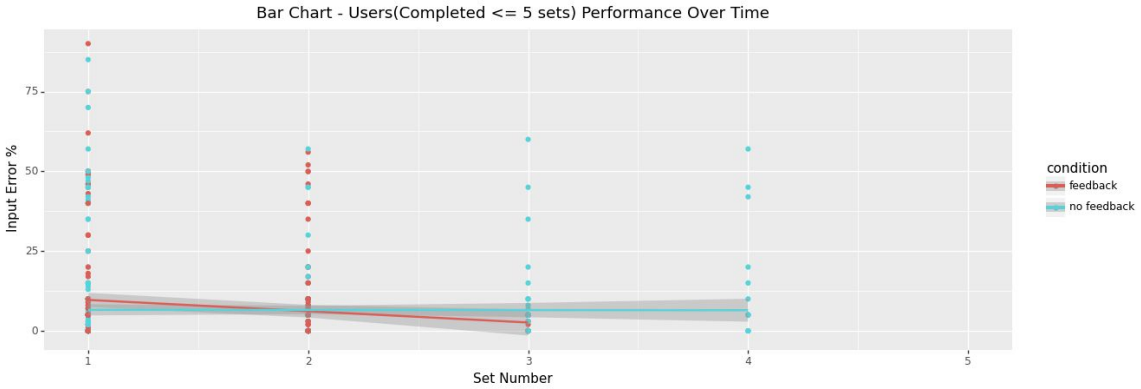

Figure 4.2.1a: Bar Chart - A scatter points diagram for participants completed less than or equal to 5 sets with local regression smooth fit.

Figure 4.2.1a shows a scatter diagram of offset of participants' judgemental error across the number of sets on bar charts. We can see that the feedback group's average judgemental error was about 11% at set one; it reduced to about 3% at set three. The no feedback group average judgemental error was about 9% at set one; it reduced to about 7% at set four.

$$(bar\ chart)\ feedback\ group\ improvement\ rate\colon \frac{11\% - 3\%}{2\ sets} = 4.5\%\ per\ set$$

$$(bar\ chart)\ no\ feedback\ group\ improvement\ rate\colon \frac{9\% - 7\%}{3\ sets} = .67\%\ per\ set$$

This data shows that the feedback group has about 3.8%% per set improvement rate faster than the no feedback group.

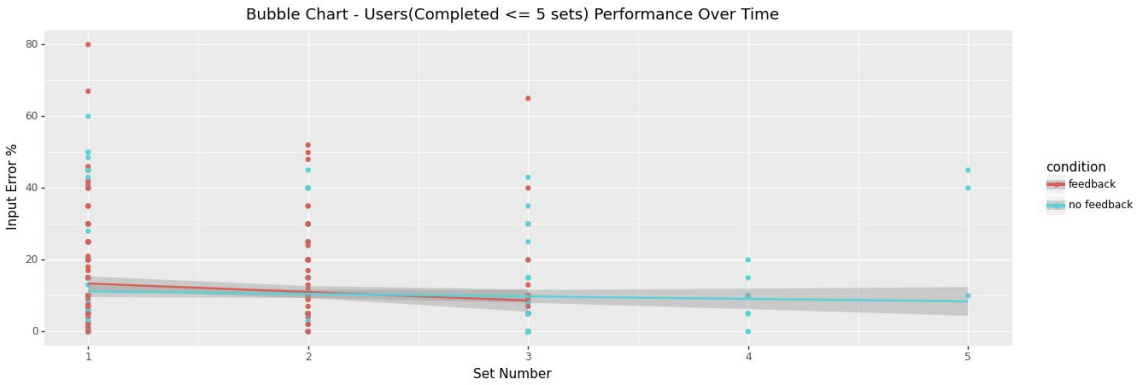

Figure 4.2.1b: Bubble Chart - A scatter points diagram for participants completed less than or equal to 5 sets with local regression smooth fit

Figure 4.2.1b shows a scatter diagram of participants' judgemental error across the number of sets on bubble charts. We can see that the feedback group's average judgemental error was about 14% at set one; it reduced to about 12% at set three. The no feedback group average judgemental error was about 13% at set one; it reduced to about 12% at set five.

$$(bubble\ chart)\ feedback\ group\ improvement\ rate\text{:}\ \frac{14\% - 12\%}{2\ sets} = 1\%\ per\ set$$

$$(bubble\ chart)\ no\ feedback\ group\ improvement\ rate\text{:}\ \frac{13\% - 12\%}{4\ sets} = .25\%\ per\ set$$

This data shows that the feedback group has about 0.8% per set improvement rate faster than the no feedback group.

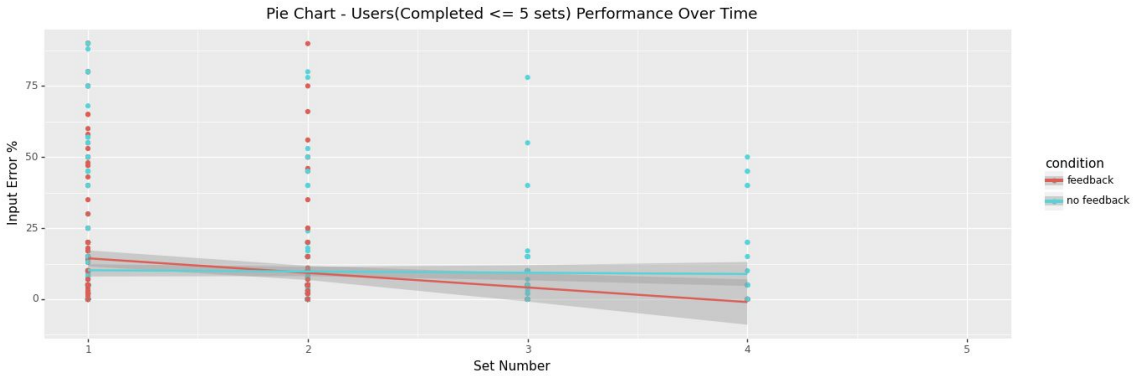

Figure 4.2.1c: PieChart - A scatter points diagram for participants completed less than or equal to 5 sets with local regression smooth fit

Figure 4.2.1c shows a scatter diagram of participants' judgemental error across the number of sets on pie charts. We can see that the feedback group's average judgemental error was about 14% at set one; it reduced to about 0% at set four. The no feedback group average judgemental error was about 12.5% at set one; it reduced to about 10% at set four.

$$(pie\ chart)\ feedback\ group\ improvement\ rate\text{:}\ \frac{14\% - 0\%}{3\ sets} = 4.7\%\ per\ set$$

$$(pie\ chart)\ no\ feedback\ group\ improvement\ rate\text{:}\ \frac{12.5\% - 10\%}{3\ sets} = .83\%\ per\ set$$

This data shows that the feedback group has about 3.9% per set improvement rate faster than the no feedback group.

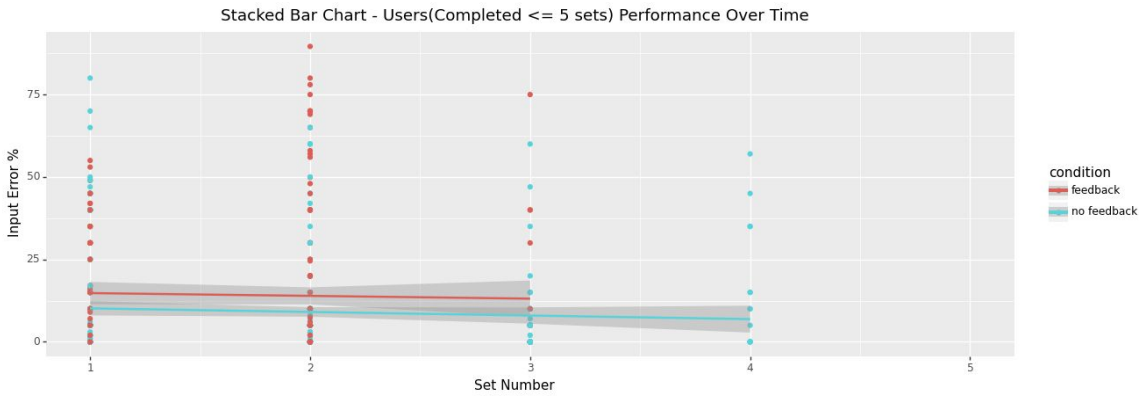

Figure 4.2.1d: Stacked Bar Chart - A scatter points diagram for participants completed less than or equal to 5 sets with local regression smooth fit.

Figure 4.2.1d shows a scatter diagram of participants' judgemental error across the number of sets on stacked bar charts. We can see that the feedback group's average judgemental error was about 14% at set one; it reduced to about 13% at set three. The no feedback group average judgemental error was about 12% at set one; it reduced to about 7% at set four.

$$(pie\ chart) feedback\ group\ improvement\ rate: \frac{14\% - 3\%}{2\ sets}. 5\%\ per\ set$$

$$(pie\ chart)\ no\ feedback\ group\ improvement\ rate: \frac{12\% - 7\%}{3\ sets} = 1.7\%\ per\ set$$

This data shows that the feedback group has about 1.2% per set improvement rate slower than the no feedback group.

4.2.2 Case Study: The Participant with 200 Trials with No Feedback

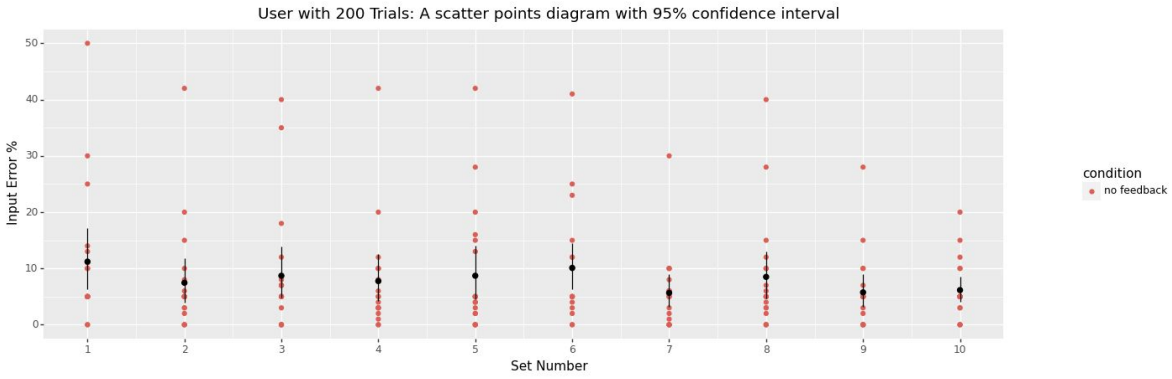

Figure 4.2.2a: A scatter points diagram for a participants who completed 200 trials with a 95% confidence interval.

Figure 4.2.2a shows a scatter diagram of judgemental error of a participant who completed 200 trials, which was the highest number of trials within all 64 participants . This particular participant was in the no feedback group. The black dot and line represents the 95 percent confidence interval. This figure shows that this participant had about a 12% average

judgemental error at the first set; the subsequent sets had lower percent judgemental error, but there was not a linear relation.

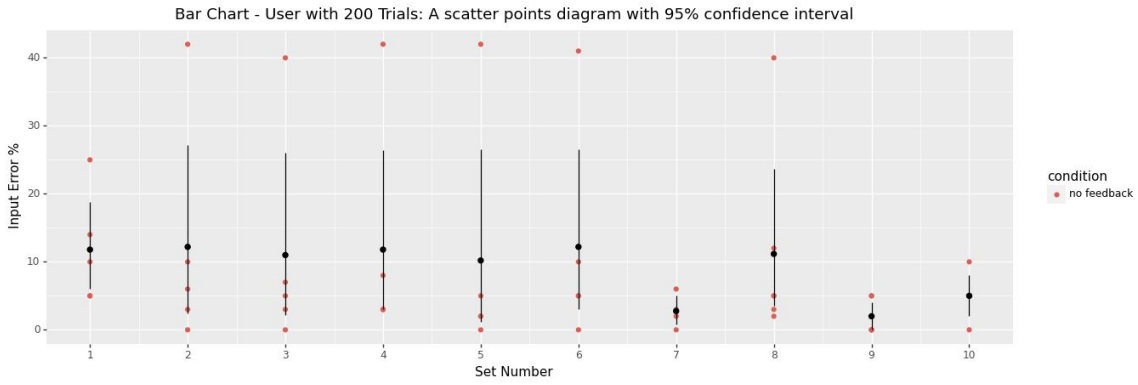

Figure 4.2.2b: Bar Chart - A scatter points diagram for a participant who completed 200 trials with a 95% confidence interval.

Figure 4.2.2b shows a scatter diagram of judgemental error of a participant who completed 200 trials on bar charts. The black dot and line represents the 95 percent confidence interval. The length of the line depicts that this data has a statistically significant difference.Although there was not a linear correlation, this figure shows that this participant's highest judgemental error was 12.5% at set two and lowest judgemental error was 4% at set nine which resulted in an overall 8.5% performance difference across seven sets.

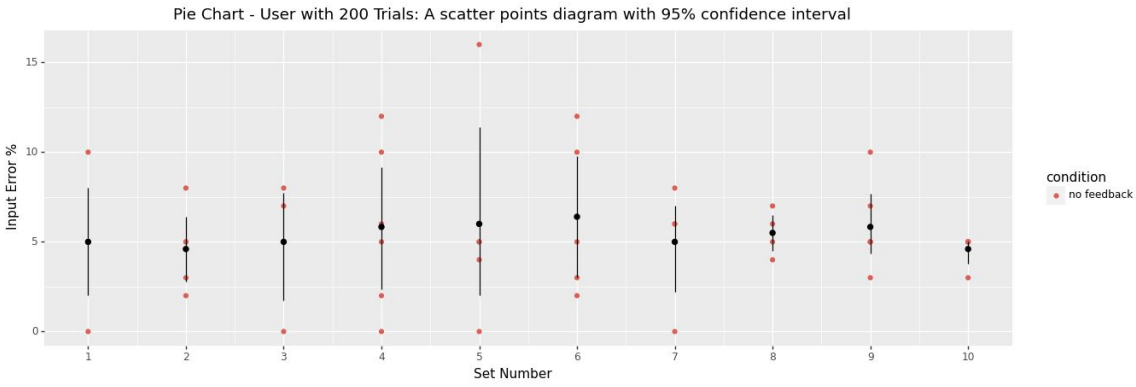

Figure 4.2.2c: Pie Chart - A scatter points diagram for a participant who completed 200 trials with a 95% confidence interval.

Figure 4.2.2c shows a scatter diagram of judgemental error of a participant who completed 200 trials on pie charts. The black dot and line represents the 95 percent confidence interval. The length of the line depicts that this data has a statistically significant difference. This shows that the participant had trouble to provide consistent input for pie charts.

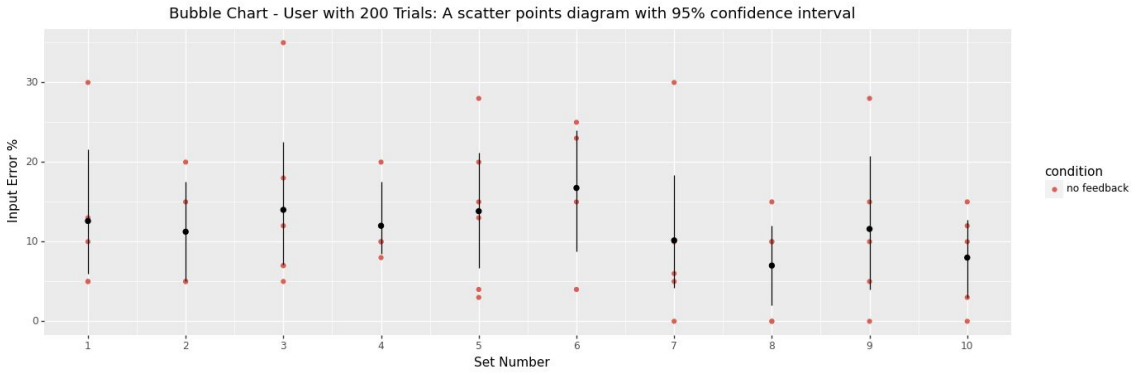

Figure 4.2.2d: Bubble Chart - A scatter points diagram for a participant who completed 200 trials with a 95% confidence interval.

Figure 4.2.2d shows a scatter diagram of judgemental error of a participant who completed 200 trials on bubble charts. The black dot and line represents the 95 percent confidence interval. The length of the line depicts that this data has a statistically significant

difference. This shows that the participant had trouble to provide consistent input for bubble charts.

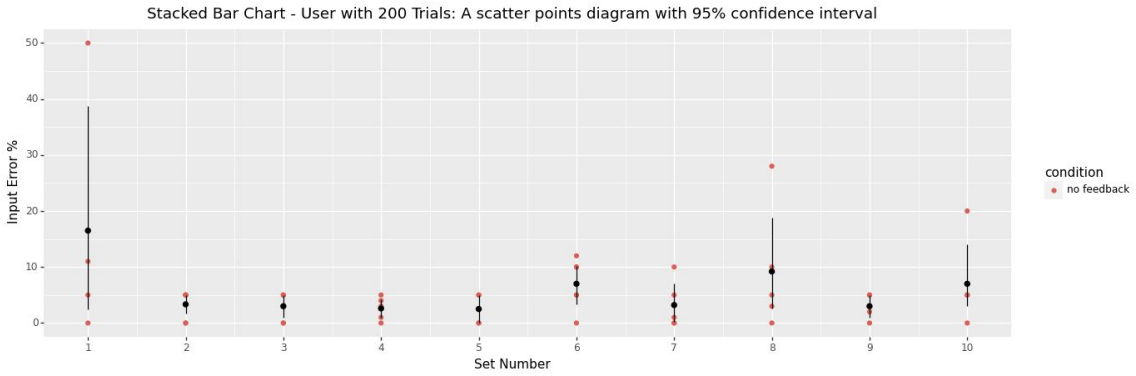

Figure 4.2.2e: Stacked Bar Chart - A scatter points diagram for a participant who completed 200 trials with a 95% confidence interval.

Figure 4.2.2e shows a scatter diagram of judgemental error of a participant who completed 200 trials on stacked bar charts. From the 95% confidence interval line, we can see this participant was highly inconsistent on the first set, but later the consistency was enhanced. This participant had a drastic improvement from the first set to the second set. However, the subsequent sets show that the participant sometimes reversed to higher judgemental error.

4.2.3 Case Study: The Participant with 140 Trials with Feedback

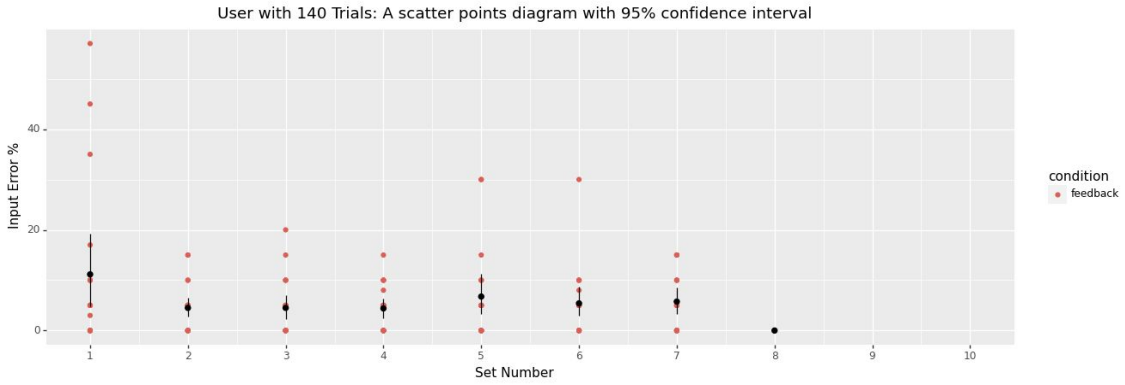

Figure 4.2.3a: A scatter points diagram for a participant who completed 140 trials with a 95% confidence interval.

Figure 4.2.3a shows a scatter diagram of judgemental error of a participant who completed 140 trials, which was the second highest number of trials within all 64 participants. This particular participant was in the feedback group. The black dot and line represents the 95 percent confidence interval. This figure shows that this participant had about a 12% average judgemental error at the first set; the subsequent sets had lower percent judgemental error, but there was not a linear relation.

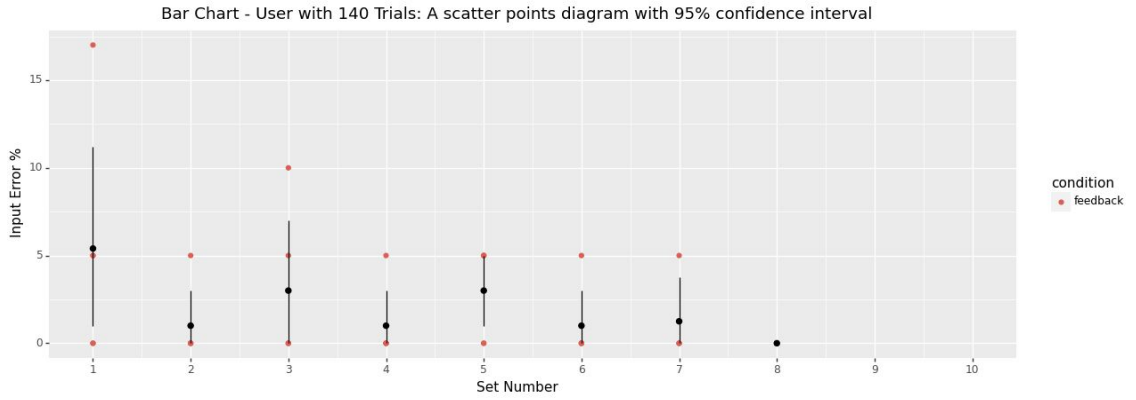

Figure 4.2.3b: Bar Chart - A scatter points diagram for a participant who completed 140 trials with a 95% confidence interval.

Figure 4.2.3b shows a scatter diagram of judgemental error of a participant who completed 140 trials on bar charts. From the 95% confidence interval line, we can see this participant was highly inconsistent on the first set, but later the consistency was enhanced. This participant had a drastic improvement from the first set to the second set. The subsequent sets had lower percent judgemental error, but there was not a linear relation.

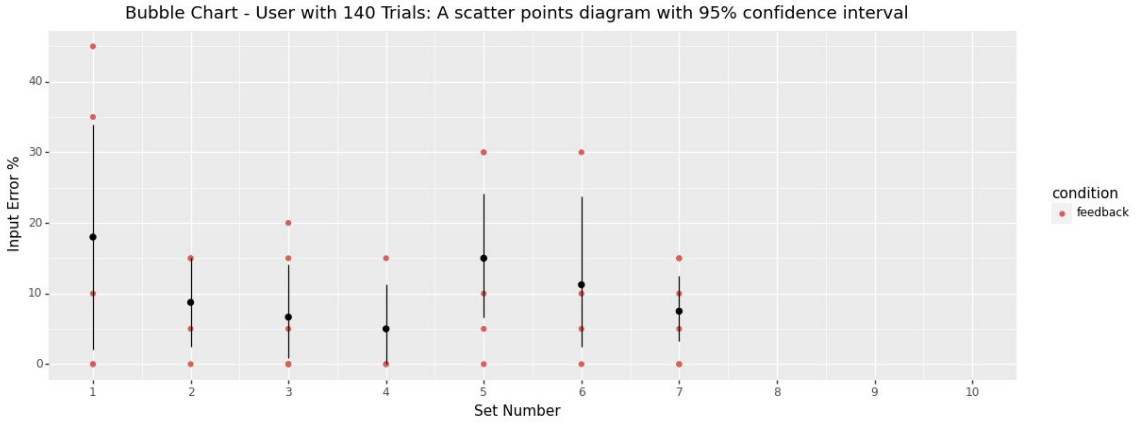

Figure 4.2.3c: Bubble Chart - A scatter points diagram for a participants who completed 140 trials with a 95% confidence interval.

Figure 4.2.3c shows a scatter diagram of judgemental error of a participant who completed 140 trials on bubble charts. From the 95% confidence interval line, we can see this participant was highly inconsistent on the first set, but later the consistency was enhanced. This participant had a drastic improvement from the first set to the second. However, the subsequent sets show that the participant sometimes reversed to higher judgemental error and higher statistically significant difference .

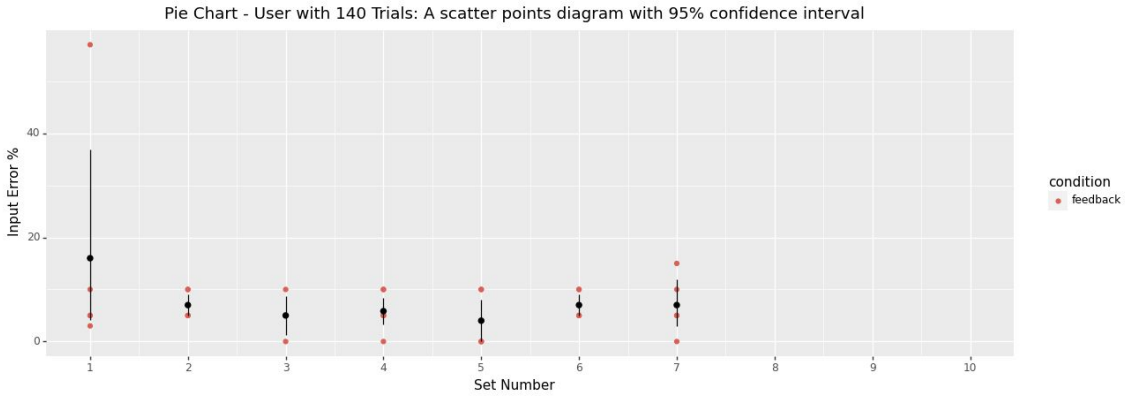

Figure 4.2.3d: Pie Chart - A scatter points diagram for a users who completed 140 trials with a 95% confidence interval.

Figure 4.2.3d shows a scatter diagram of judgemental error of a participant who completed 140 trials on pie charts. From the 95% confidence interval line, we can see this participant was highly inconsistent on the first set, but later the consistency was enhanced. This participant improved from 17% judgemental error to 4% judgemental error over the first five sets.

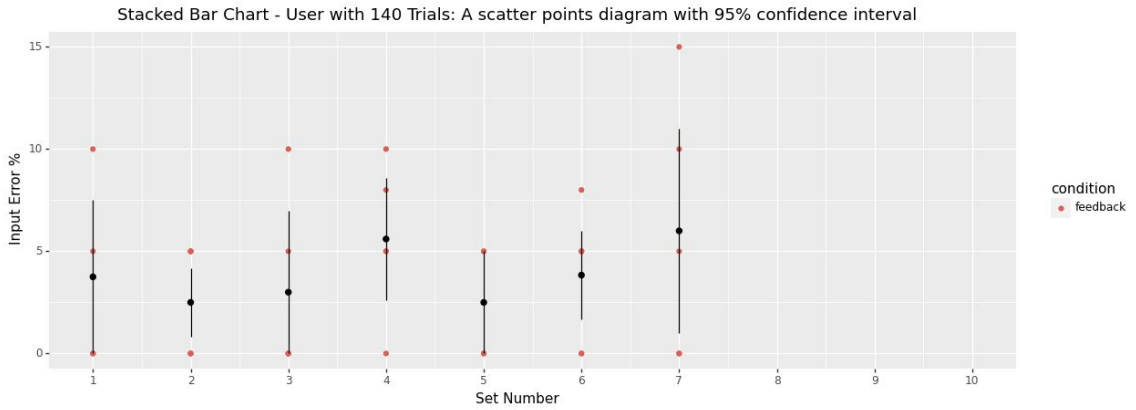

Figure 4.2.3e: Stacked Bar Chart - A scatter points diagram for a participant who completed 140 trials with a 95% confidence interval.

Figure 4.2.3e shows a scatter diagram of judgemental error of a participant who completed 140 trials on stacked bar charts. The black dot and line represents the 95 percent confidence interval. The length of the line depicts that this data has a statistically significant difference. This shows that the participant had trouble to provide consistent input for stacked bar charts.

Chapter 5 Conclusion

Data visualization has become increasingly popular in the past decades in all aspects of our lives, from assisting us in representing data during a presentation to visualizing traffic flow and weather forecast. With online crowdsourcing platform growth, many researchers choose to conduct online studies instead of lab experiments. Over the past decades, data visualization assessment has proven a quantitative hypothesis while changing its platform from lab experiment to online crowdsourced studies. However, we noticed that these studies did not consider applying feedback which is a missed opportunity to grasp the effect of feedback in data visualization assessment.

This report gathered feedback mechanics from educational platforms, video games, and fitness applications and summarized some design principles for feedback: inviting, repeatability, coherence, and data driven. We conducted an online crowdsourced data visualization experiment on Prolific.co. The data we gathered show that in average 1) the feedback group learns accurate data visualization reading slightly faster, 2) the no feedback group completed more sets than the feedback group, and 3) the no feedback group has better performance on the first set. While the first result meets our expectation which sort of verifies the effect of feedback, the two succeeding results are complicated and need more comprehensive studies to explain them. For future work, we would like to conduct a large scale online research to affirm that feedback can improve participants' data visualization reading performance and unravel some uncertain trends we saw in the data which could be influenced by feedback conditions, demographic criteria, and etc.

Reference:

[1] Aubin, Casey. The Importance of Immediate Feedback in Learning. Online at: https://www.smartickmethod.com/blog/education/pedagogy/inmediate-feedback/

[2] Beger, Rudolf. Present-Day Corporate Communication: A Practice-Oriented, State-of-the-Art Guide. 2018.

[3] Clarà, Mare; Barberà, Elena. Learning online: massive open online courses (MOOCs), connectivism, and cultural psychology, 2013. DOI: 10.1080/01587919.2013.770428. Online at: https://www.tandfonline.com/doi/abs/10.1080/01587919.2013.770428

[4] Cleveland, William S.; McGill Robert. Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *Journal of the American Statistical Association, Vol 79*, 1984.

[5] Downes, Stephen. Connectivism' and Connective Knowledge, *Huffpost Education*, 2011.

[6] Few, Stephen. Tapping the Power of Visual Perception. *Perceptual Edge*, 2004.

[7] Gigaom. Duolingo snags iPhone App of the Year, 2013.

[8] Harrison, Lane; Skau, Drew; Steven, Franconeri; Lu, Aidong; Remco, Chang. Influencing Visual Judgment through Affective Priming, 2013.

[9] Harrison, Lane; Yang, Fumeng; Franconeri, Steven; Chang, Remco. Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design, 2014.

[10] Heer, Jeffrey; Bostock, Miachel. Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design, 2010.

[11] Kaplan, R.; Kaplan, S. The Experience of Nature: A Psychological Perspective. 1989.

[12] Kapp, Karl. Feedback Essential for Video Games and Learning . Online at : http://karlkapp.com/feedback-essential-for-video-games-and-learning/ [Kapp]

[13] Knuth. Marc. The Critical Role of Data Visualization in Better Healthcare. January 16, 2013. Online at: https://www.mckesson.com/blog/data-visualization-role-future-of-healthcare/

[14]Maslow, A.H.. A theory of human motivation. *Psychological Review* 1943. DOI:10.1037/h0054346

[15] Miller, G. A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review 63*, 1956. doi:10.1037/h0043158

[16] Mistra, Satish. New report finds more than 165,000 mobile health apps now available, takes close look at characteristics & use, 2015. Online at: https://www.imedicalapps.com/2015/09/ims-health-apps-report/

[17] P. M. Inness and S. Dorling, Operational Weather Forecasting. *Wiley-Blackwell*, 2013.

[18]Reinecke, Katharina; Guo, Philip J. Demographic Differences in How Students Navigate Through MOOCs. DOI: http://dx.doi.org/10.1145/2556325.2566247

[19] Schaefer, Charles; Millman, Howard. How to Help Children with Common Problems. 1994.

[20] Weber's Law definition. *USD Internet Sensation & Perception Laboratory.* Online at: http://apps.usd.edu/coglab/WebersLaw.html

[21] Wilson, Karen, Korn, James H. (5 June 2007). Attention During Lectures: Beyond Ten Minutes. *Teaching of Psychology. 34 (2): 85–89,* 5 June 2007.

Appendix A - Consent Form

Informed Consent Agreement for Participation in a Research Study

Investigator: Lane Harrison, Yiren Ding, and Meixintong Zha

Contact Information: ltharrison@wpi.edu; 980-200-8363

yding5@wpi.edu;

mxzha@wpi.edu;

Title of Research Study: Validating Model-based Approaches for Data Visualization Ability Assessment

Introduction

You are being asked to participate in a research study. Before you agree, however, you must be fully informed about the purpose of the study, the procedures to be followed, and any benefits, risks or discomfort that you may experience as a result of your participation. This form presents information about the study so that you may make a fully informed decision regarding your participation.

Purpose of the study:

Many news organizations and companies are producing visualizations to communicate complex data in novel and engaging ways. The purpose of this study is to examine techniques that help people interpret data visualizations and model how well people can perform these tasks.

Procedures to be followed:

You will be shown a series of data visualization. We'll ask you answer basic questions about the chart and the data it contains. You'll have several warm-up trials at the beginning, the correct answer will be given after you submit your answer. In addition, you may need to fill out demographic survey, spatial ability survey or visualization literacy assessment test.

Risks to study participants:

Each participant in this study is assigned a random ID. As such, your participation will remain anonymous and your responses will not able to be used to identify you.

Benefits to research participants and others:

The possible benefits include exposure to interesting data visualization techniques and topics, along with helping inform the development of future data visualization techniques.

Record keeping and confidentiality:

Records of your participation in this study will be held confidential so far as permitted by law. However, the study investigators, the sponsor or it's designee and, under certain circumstances, the Worcester Polytechnic Institute Institutional Review Board (WPI IRB) will be able to inspect and have access to this data. Any publication or presentation of the data will not identify you.

Cost/Payment: $1.25 (This study is estimated to take approximately 10 minutes.)

For more information about this research or about the rights of research participants, or in case of research-related injury, contact:

Lane Harrison (contact info at the top of this page). In addition, include the contact information for the IRB Manager (Ruth McKeogh, Tel. 508 831- 6699, Email: irb@wpi.edu) and the Human Protection Administrator (Gabriel Johnson, Tel. 508-831-4989, Email: gjohnson@wpi.edu).

Your participation in this research is voluntary.

Your refusal to participate will not result in any penalty to you or any loss of benefits to which you may otherwise be entitled. You may decide to stop participating in the research at any time without penalty or loss of other benefits. The project investigators retain the right to cancel or postpone the experimental procedures at any time they see fit.

By clicking below,

you acknowledge that you have been informed about and consent to be a participant in the study described above. Make sure that your questions are answered to your satisfaction before signing. You are entitled to retain a copy of this consent agreement(print).

Appendix B - Survey for the feedback group

Submission

Thank you! Before you submit, please fill out the following survey.

1. In your opinion, what types of feedback on your chart reading performance, if provided, would improve your learning experience? *

2. Your gender: *

Female Male Non-binary/third gender

Prefer to self-describe Prefer not to say

3. Your age: *

4. Your country of origin: * Please select a country.

5. Highest degree obtained: *

High School Associates Bachelors Masters

6. Please indicate how experienced you are with data visualization concepts and tools. (1 means "I have never used data visualization" and 7 means "I use data visualization in my daily work") *

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7

7. Please indicate how experienced you are with statistics. (1 means "I have never used statistics" and 7 means "I use and analyze statistics in my daily work" *

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7

8. Your monitor size is closest to: *

○ 9" or smaller ○ 11" ○ 13" ○ 15" ○ 17" ○ 19" ○ 21"

○ 23" or larger ○ I'm not sure.

9. Out of the tree types of feedback, which one helps you the most? *

- ◯ (1) the per-trial summary of error (with icons for chart type)
- ◯ (2) the aggregate bar chart of error
- ◯ (3) the donut charts of error per chart type

10. Out of the tree types of feedback, which one helps you the least? *

- ◯ (1) the per-trial summary of error (with icons for chart type)
- ◯ (2) the aggregate bar chart of error
- ◯ (3) the donut charts of error per chart type

11. Do you think the feedback provided is helpful for you to reach 80% correct on all charts? Briefly explain your answer. *

12. Please share anything else you'd like to tell us here:

Appendix C - Survey for the no feedback group

Thank you! Before you submit, please fill out the following survey.

1. In your opinion, what types of feedback on your chart reading performance, if provided, would improve your learning experience? *

2. Your gender: *

Female Male Non-binary/third gender

Prefer to self-describe Prefer not to say

3. Your age: *

4. Your country of origin: * Please select a country.

5. Highest degree obtained: *

High School Associates Bachelors Masters

6. Please indicate how experienced you are with data visualization concepts and tools. (1 means "I have never used data visualization" and 7 means "I use data visualization in my daily work") *

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7

7. Please indicate how experienced you are with statistics. (1 means "I have never used statistics" and 7 means "I use and analyze statistics in my daily work" *

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7

8. Your monitor size is closest to: *

○ 9" or smaller ○ 11" ○ 13" ○ 15" ○ 17" ○ 19" ○ 21"

○ 23" or larger ○ I'm not sure.

9. Please share anything else you'd like to tell us here: