October 2021

Comparing and Improving the Design of Physical Activity Data Visualizations

Peter M. Frackleton
University of Massachusetts Amherst

COMPARING AND IMPROVING THE DESIGN OF PHYSICAL ACTIVITY DATA VISUALIZATIONS

A Thesis Presented

by

PETER FRACKLETON

Submitted to the Graduate School of the

University of Massachusetts Amherst in partial fulfillment

of the requirements for the degree of

MASTER OF SCIENCE IN

INDUSTRIAL ENGINEERING AND OPERATIONS RESEARCH

September 2021

Mechanical and Industrial Engineering

COMPARING AND IMPROVING THE DESIGN OF PHYSICAL ACTIVITY DATA VISUALIZATIONS

A Thesis Presented

by

PETER FRACKLETON

Approved as to style and content by:

_____
Jenna Marquard, Chair

_____
Shannon Roberts, Member

_____
Cynthia Jacelon, Member

_____
Sundar Krishnamurty, Department Head
Mechanical and Industrial Engineering

ACKNOWLEDGEMENTS

I would like to thank my thesis advisor, Dr. Jenna Marquard, for her guidance and support in each phase of this thesis. I would also like to acknowledge thesis committee members Dr. Cynthia Jacelon and Dr. Shannon Roberts for their involvement in and support of this thesis. Lastly, I would like to thank Rare Patient Voice (https://rarepatientvoice.com) for their assistance in reaching out to potential candidates for participation in this thesis study.

ABSTRACT

COMPARING AND IMPROVING THE DESIGN OF PHYSICAL ACTIVITY DATA VISUALIZATIONS

SEPTEMBER 2021

PETER FRACKLETON, B.S., UNIVERSITY OF MASSACHUSETTS AMHERST, M.S., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Jenna Marquard

Heart disease is a leading cause of death in the United States, and older adults are at highest risk of being diagnosed with heart disease. Consistent physical exercise is an effective means of deterring onset of heart disease, and physical activity tracking devices can inspire greater activity in older adults. However, physical activity tracking device abandonment is quite common due to limitations on what can be learned from the activity data that is collected. Better data visualization of physical data presents an opportunity to surpass these limitations. In this thesis, a task-based human subject study was performed with three different data visualizations to gain insight into how the format of physical activity data visualizations impact older adults' abilities to infer meaning from physical activity data. Participants (n = 30) interacted with a prototype data visualization as well as two data visualizations from popular fitness tracking applications (Fitbit and Strava) and used these visualizations to complete 11 tasks. Results from these tasks show each visualization was able to facilitate users answer some task questions effectively, though no visualizations exhibited strong performance across all tasks. From the successes and shortcomings of each visualization, three key design recommendations for the design of data visualizations for physical activity data were made: 1) make exact values available, 2) summarize data at multiple timescales, and 3) ensure accessibility for the entire population of users.

TABLE OF CONTENTS

APPENDICES

LIST OF TABLES

LIST OF FIGURES

Figure **Page**

CHAPTER 1

INTRODUCTION

According to a 2020 report from the American Centers for Disease Control and Prevention (CDC), an estimated 655,000 people die from heart disease per year, making heart disease a leading cause of death within the United States across all demographics [1]. Heart disease is an especially formidable threat to those aged 65 and older, who account for roughly 8 in 10 of deaths due to heart coronary heart disease [1,2]. Though heart disease is prevalent among adults, studies have consistently shown throughout the past several decades that regular exercise to improve cardiorespiratory fitness can effectively deter onset of cardiovascular disease as well as rehabilitate individuals who have experienced heart failures [3-5]. However, the CDC's Behavioral Risk Factor Surveillance System reported that in 2019, only approximately 50% of adults in the United States were performing the recommended amount of physical activity per week across all age groups [6]. With older adults being the part of the population with highest risk of succumbing to heart disease [1,2], increasing the number of older adults who perform the recommended amount of physical activity per week presents an opportunity to greatly reduce the number of deaths to heart disease each year.

Wearable physical activity tracking devices have been observed to possess the potential to motivate higher levels of physical activity in users [7]. In the case of older adults specifically, multiple longitudinal studies of the effects of tracking technology interventions that ranged in length from eight weeks to six months weight loss showed improved adherence of nutrition plans [8] and increased energy expenditure through physical activity [9-11] when participants used tracking technology to manage their exercise or nutrition plan. However, this motivation and the overall use of wearable

devices does not always persist over long periods of time, as explored in two studies by the DUB group from the University of Washington [12,13]. Each of these studies analyzed survey responses from over 100 participants to establish patterns in their reasons for lapsing in or abandoning tracking. Two of these reasons for lapsing in or abandoning tracking that were present in both studies are that, over time, users found themselves to be learning nothing new about their behaviors and there was no clear feedback from their data on what habits to change or how [8,9]. In the case of older adults specifically, a similar phenomenon occurred in an eight-month trial, during which participants used a commercially available monitor to keep track of physical activity. Over this eight-month trial, older adults generally rated this monitor as less easy to use and less useful than they had at the beginning of the study [14]. Research intro improving engagement with fitness technology has been performed, exploring different behavior change strategies such as goal setting, self-monitoring, and feedback, all of which have been shown to increase activity and healthy behaviors [15]. Data visualization has the potential to provide information that is critical to all three of these strategies. Though basic data visualization is already present to many users of physical activity tracking devices and apps directly on the device or in the device's app, it is unclear how these data visualizations are used and what information they can provide to users in the context of insights into physical activity behavior and overall health. By evaluating the strengths and limitations of current data visualizations and using this information to create better data visualizations that provide users with more information, there is potential to improve long-term adoption of this technology and promote more frequent physical activity.

Currently, no standards exist for best designing these types of physical activity data visualizations so that they consistently provide users with useful and usable information. The focus of this thesis is to understand what features within three physical

activity data visualizations allow a user to draw meaningful insights about physical activity behaviors using device-generated data, and which design elements make data visualizations easier or more difficult to interpret. The research questions to be answered in this thesis are:

1. How does the format of physical activity data visualizations impact older adults' abilities to infer meaning from physical activity data? Ability to interpret meaning is broken down into two key factors:
    a. Task response correctness
    b. Task completion time
2. Are there other variations among users that impact an older adult's ability to infer accurate meaning from physical activity data? Variations considered include:
    a. Demographics (level of education, career field, gender)
    b. Current level of physical activity
    c. Level of experience with fitness tracker technology
    d. Graphical literacy

These research questions were answered through the collection and analysis of data acquired from 30 participants aged 55 and above. Participants interacted with three data visualizations displaying physical activity data: two on-market data visualizations and one prototype data visualization, and then answered eleven task questions using each data visualization. Participants were then surveyed and interviewed about their experiences performing these tasks to understand why some tasks were easy or difficult for them to complete. Next, participants were interviewed about which data visualization was most preferable to them and what was most important to them if tracking their physical activity. Lastly, data for additional characteristics such as prior experience with

fitness tracking apps, graphical literacy, and physical activity level were collected through questionnaires.

The contents of this study provide three major contributions to data visualization design for physical activity tracking: 1) evidence that current data visualization design for popular physical activity tracking applications (Fitbit, Strava) do not present physical activity data in a way that allows an older adult to quickly and accurately answer all basic questions about the contents of their data, 2) an evaluation of how informative and usable Fitbit and Strava's data visualizations and how they compare to the evaluation of a physical activity data visualization prototype, and 3) recommendations for how to design a more informative and usable data visualization for physical activity based on the performance of each visualization for answering various basic questions about physical activity data.

This thesis is structured as follows. Chapter 2 summarizes findings from related studies. Chapter 3 describes the study methods and provides detail on the rationale for the study design. Chapter 4 presents and explains the study results. Chapter 5 discusses the implications of the findings presented in Chapter 4, including how the findings can inform the design of physical activity data visualizations. Chapter 6 summarizes the purpose and key findings of this study.

CHAPTER 2

LITERATURE REVIEW

The research in this thesis is centered around data visualization of physical activity data. To better understand this application of data visualization, literature from a wide range of topics was reviewed to create a foundation for this thesis and its contribution. The topics that were reviewed for this thesis study include 1) the importance of electronic health data, 2) what individuals are learning from their health data and data visualization's role in this process, 3) what qualities an effective data visualization has, 4) how data visualization effectiveness may vary with different users, and 5) methods of evaluating if a data visualization is effective or not. A separate section is dedicated for each of these topics where key reviewed literature for these findings is described.

2.1 Electronically Capturing and Tracking Patient Data

Over the past decade, there has been a growing emphasis on health information technology (HIT), referring to technologies designed to collect and use health data and knowledge for healthcare-related communication and decision making [16]. As part of part of the American Recovery and Reinvestment Act of 2009 (ARRA), the Health Information Technology for Economic and Clinical Health (HITECH) Act was passed, allocating $19 billion to promote the adoption and meaningful use of HIT [17,18], with meaningful use referring to applications of HIT that would have a positive impact on quality of care, as opposed to HIT being applied to scenarios where there is no clear benefit [19]. The Office of the National Coordinator (ONC) now provides certification for electronic health records (EHRs) that meet requirements under what is now known as the Promoting Interoperability Programs [20]. One of the requirements for EHR

certification from the ONC that pertains to meaningful use is providing patients electronic access to their health information [20]. Many certified EHRs offer patients electronic access to their health information through patient portals, for the sake of keeping their patients better informed about their health and improving quality of care. These portals include information such as medications being taken, lab results, and a list of known medical problems [21]. As access to patient portals becomes more commonplace, more patients are accessing electronic health information collected during visits to healthcare providers, ideally using the information as a tool to improve their health [22].

HIT is not limited only to information collected in clinical settings. Some clinicians are also incorporating patient-generated health data (PGHD) collected outside of clinical settings into their EHRs [23,24]. This integration can lead to clinicians having deeper insight into a patient's health between visits, and patients may be able to gain health insights to become more empowered to manage their own health [25,26]. Examples of the types of PGHD that clinicians and patients can benefit from ranges from apps requiring manual entry of data to data automatically collected via mobile sensors. A common form of PGHD that is central to this thesis is the data that comes from wearable fitness trackers made by device manufacturers such as Fitbit, Samsung, and Apple. These wearable fitness trackers are widely used, and the user pool continues to grow [27]. These devices can automatically collect physical activity data from the user, such as distance walked or steps taken within a time frame. These data are then available on the wearable device itself, a phone app, or website in the form of raw data or data visualizations for the user to examine and reflect upon. Integrating these PGHD into the EHRs could supplement clinical EHR data, linking data about activities of daily living (ADLs) with clinical status and outcomes and providing a more holistic view of the status of one's health.

2.2 Use of Tracking Technology

With the commercial growth of devices and apps that can collect PGHD, practices of tracking one's own personal health data are becoming increasingly popular outside of clinical settings as well. One study reported that 69% of US adults have engaged in personal tracking of health or other data [28]. This practice is commonly referred to as self-quantification or personal informatics, and the data enthusiasts who engage in it have been dubbed by some as Quantified-Selfers [29, 30]. Though practices in self-quantification are varied, Li et al. interviewed fifteen individuals using different types of personal data to establish a taxonomy for what types of questions users were trying to answer with their personal data. The resulting taxonomy contains six categories, which are: ***status*** (information about current progress such as the number of steps walked so far that day); ***history*** (refers to seeking trends and patterns over longer periods of time); ***goals*** (developing short or long-term objectives such as number of times to exercise in a week); ***discrepancies*** (ex: the difference between status and goal at a given time); ***context*** (describing other discrete events that may have influenced measurement); and ***factors*** (how one measure influences another such as how nutrition and physical activity affect overall health). It was also suggested that these questions can occur at one of two phases: ***discovery***, which is concerned with learning what goal they are trying to meet or identifying relationship between factors, or ***maintenance***, in which an individual primarily maintains awareness of their status or maintains a behavior [31].

A useful tool within personal informatics to facilitate answering such questions is data visualization, as well-designed data visualizations allow users to quickly explore large amounts of data observe important insights from large amounts of data. When designed correctly and useful to users, data visualizations provide an opportunity to keep users more informed and engaged in tracking behaviors [32]. However, despite the

value of the data collected by these users, it is common for these users to lapse in this type of tracking, either temporarily or permanently. Some users lapse in tracking due to forgetting to charge their devices frequently or simply losing interest in tracking their data. Other users may lapse in tracking because of the frustration they experience when they find that the tracking tools that they use do not help them to reach a goal or gain actionable insight, or when they find the data too difficult to understand [13]. When some highly dedicated individuals determine that the default visualizations provided within fitness tracking apps are not considered adequate, they seek out an alternative tool to visualize their data and some even go so far as to make their own visualizations of their data using tools such as d3 or Google Charts [29]. However, most of these Quantified-Selfers do not have expertise in visualization design, so these visualizations are not as useful as intended [33]. Data visualization experts therefore need to understand what end users seek in their data and design visualizations with these considerations in mind.

2.3 Data Visualization Design

Creating an effective data visualization tool for individuals who are tracking their personal data relies on utilizing data visualization design principles effectively. Evergreen (2017) puts forward that the starting point of a data visualization is identifying the point that this visualization is trying to make, because every design choice made in a data visualization will influence how easy or difficult it is for that point to be communicated [34]. There are a wide variety of key choices to make when designing a data visualization, such as the method of visual encoding, layout, color palette, aesthetic styling, and interactions. The number of possibilities for each of these aspects of design and how these decisions are made are vast, but some basic concepts for each remain consistent as fundamentally good design, and within the scope of representing historical

exercise data, some more specific studies that have been performed in the past that contribute useful feedback from users.

Because the practice of self-quantification involves the collection and review of self-generated data, many physical activity tracking apps such as Apple Health, Fitbit, Samsung Health, and Strava visualize past physical activity as a time series. Common visual encoding methods for time series data are bar, line, dot, and dot bar graphs. A standard cartesian layout in which a horizontal axis and vertical axis are used to represent progression of time and magnitude of value for the visually encoded data in most cases, but data layouts for this application of data can expand to include Gregorian calendars as well as radial plots, especially when a regular cycle is being visualized [35]. Which of the methods of encoding should be chosen will depend on what type of time series relationship is the focus of the visualization [36].

Shifting focus to layout, color palette, and general aesthetic styling, adopting concepts from HMI design for machine interfaces would suggest that factors such as color, font, or layout can positively or negatively impact usability [37]. Vibrant background color, irregular font, and unintuitive layout of visualization elements can all lead to poor readability frustrate the user [37]. Aesthetic appeal can also impact performance even when the graph is still perfectly readable. A study collecting responses from 285 participants asked participants to rate the perceived aesthetic appeal of seven visualizations individually, rank them with respect to each other, and then perform tasks with them [38]. Each of the visualizations employed the same color palette, size and scaling, typography, and data. Results showed that participants performed tasks most effectively with visualizations that they found visually appealing, having fewer incorrect responses to tasks and faster task completion time [38].

Though much of the design of a visualization can be guided by considering the message that is intended to be communicated, the medium through which a visualization

is being viewed is also important to design around since some mediums may have more limitations than others. Apple Health, Fitbit, Samsung Health, and Strava, some of the most popular apps that are centered around tracking, visualize data on mobile devices. Mobile devices have different user input methods, smaller screen sizes, and different screen ratios compared to a traditional desktop computer. Chittaro (2006) outlines a set of six major steps that should be addressed when designing visualizations for a mobile device. One of these steps is to identify what available tools to help a user navigate their data. Since not as much data can be displayed on the screen clearly, mobile visualizations typically include tools such as filters and zooming to allow the user to get a more detailed view of the data that they are interested in [39]. Games (2014) approached the limited screen space by implementing various features and interactions such as zooming, displaying data with a fisheye view, and borders that would indicate in which direction off-screen data was concentrated. That study concluded that the addition of the off-screen contextual information was helpful to participants in identifying the data in the tasks they were given, showing that this design has some degree of validity and could be a method to overcome the challenges posed by the small screens on mobile devices [40].

Though data visualizations for fitness tracker data tend to use more traditional forms of encoding such as bars to represent time series data on a linear axis, new concepts are also being explored. Ambient data visualizations, sometimes referred to as informative art, have been considered as a method to increase user awareness of their physical activity data through visually appealing and easily accessible data visualizations [41]. Fan (2012) examined how users would interact with simple, colorful ambient visualizations of a user's physical activity placed in locations that participants regularly visited [41]. In this study, the immediate availability and visual allure of these visualizations was shown to increase awareness of current physical activity status

compared to traditional fitness tracker interfaces. However, despite the appeal of these visualizations, participants preferred more traditional graphical visualizations when looking for specific information or historic patterns [41].

Although it is important to identify what a data visualization intends to communicate before creating it [34], as seen in the taxonomy developed by Li et al. (2011) [31], users of tracking technology may have many different types of questions about their data that they would like to answer. Creating a single data visualization to answer all questions that all users could have would be unlikely to allow a user to answer all their questions efficiently or easily as the visualization is liable to end up lacking information or busy and overloaded with information. Choe et al. (2017) proposes a solution in the form of a semi-flexible systems where users can explore their own data. Researchers developed and orchestrated an in-lab think aloud study on a web-based application called Visualized Self that allowed individuals to visualize their personal data in different ways to understand how it could assist users in gaining richer insights from their data [42]. Two key unique features of Visualized Self were its ability to incorporate contextual information to data and flexibility in how data was visualized. By default, context of activity that occurred on weekends was visualized, but participants had the ability to add other context to data. In general, participants appreciated the flexibility of the system that allowed them to choose between different visualizations of the data and were able to use the system to test hypotheses about their data as they incorporated context into data. However, participants also expressed that they would prefer a system that would incorporate more sophisticated methods of contextual information into their data, such as visualizing work hours and when activity occurs throughout the day [42].

Two studies taking a similar approach to incorporation of contextual information into data collection have also been performed. For reviewing sleep data, visually

incorporating contextual information into the data may help a user establish the causality between poor sleep quality and other behaviors or understand which of their sleeping patterns lead to feeling well-rested. Liang et al. (2016) focused on tracking sleep data and allowing for incorporation of contextual factors such as electronic device usage or caffeine intake. The field study that followed 12 participants showed positive results, with users feeling both more informed than when they had just sleep data alone and able to make behavior changes based on the information obtained [43]. Pavel et al. (2013) incorporated contextual information into data through a "story-inspired paradigm," including information such as location, people, and theme and affiliating it with data collected by a user, finding that users felt more engaged with this form of data visualization [44]. Though none of these studies determined whether users were able to gain greater insight into their data in a quantifiable manner, higher user satisfaction and engagement seemed to be a consistent outcome of participant interaction with these prototypes.

2.4 Variation Among Users

As established in the previous section, it is important to know what types of questions the audience for whom you design data visualizations intends to answer, as these questions will shape *what* the data visualization needs to communicate. However, it is also important to recognize how well your audience can read a data visualization, because this places limitations on *how* the data visualization can communicate the intended message to the audience. Graphical literacy refers to how well an individual can understand a graphical representation of data [45]. Someone with low graphical literacy is more likely to examine the wrong parts of or misread a graph more often than someone with high graphical literacy [46]. Graphical literacy cannot be reliably inferred based upon a person's education [47], age [47] or race [47,48], and graphical literacy

has not been found to vary based solely on geographical location or nationality [49]. Since no validated predictors of graphical literacy exist, tests to assess an individual's level of graphical literacy have been created. One such test is the Graphical Literacy Scale (GLS), which has been validated [50] and is the instruments chosen to measure graphical literacy in the experiment conducted for this thesis.

Another external factor that may affect an individual's ability to understand a visualization is their level of expertise on the subject matter being visualized. For example, in the domain of business, a novice may draw fewer inferences looking at a business-related graph than an expert. However, this phenomenon has only been observed to be true in some domains, and to varying degrees across those domains [51]. To address the possibility of this effect on viewing data visualizations for physical activity data, participants of this study completed two questionnaires: the Simple Physical Activity Questionnaire (SIMPAQ) as a validated tool to assess their level of physical activity [52], and the fitness tracking experience questionnaire which was developed specifically for this study to quantify a participant's familiarity with physical activity tracking devices.

Aside from the influence of graphical literacy and subject knowledge, data visualization usability can also be limited for individuals because of accessibility issues. One of the most common issues that is considered is color blindness, as this affects what color palette can be used for the data visualization, and which ones will be visually discernable for all types of color blindness. Multiple free tools exist online [53], generally referred to as a "color blindness simulator" or similar, through which you can upload an image to see how it may be visible to individuals with different forms of color blindness and adjust a design or color palette accordingly. Another accessibility concern that is relevant to this study is the reduced visual contrast sensitivity in older adults. Those aged 60 or above have been found to have significantly lower contrast sensitivity than

younger individuals for middle and high spatial frequencies [54]. Use of strong color contrast and sharp lines are important to designing a readable data visualization, especially for older adults.

2.5 Evaluating Data Visualizations

Although a data visualization's design can be driven by focusing on answering specific questions and considering how the intended audience will view, interpret, and use the information presented, some form of evaluating the design is necessary to validate it. Two forms of evaluation are reviewed in this section: heuristic evaluation and evaluation through controlled user studies. A well-known set of heuristics in the field of human-computer interaction is Nielsen's 10 usability heuristics for user interface design, which lists 10 rules for design meant to outline the characteristics of an effective interface [55]. Forsell et al. (2010) formulated 10 heuristics for evaluation of information visualization that serves a similar function as Nielsen's 10 usability heuristics, but for the field of data visualization [56]. Some of the heuristics included in this list are minimal actions, referring to requiring few actions from the user to accomplish a task; consistency, referring to maintaining similar design in similar context and different design in different context; spatial organization, referring to clear and intuitive layout with efficient management of space; and data set reduction, referring to only showing what part of the data set is necessary to complete a task efficiently [56]. Though the this set of heuristics was chosen to provide the widest coverage of visualization design considerations, these heuristics are not claimed as a final set. A subsequent study recommended addition of heuristics related to interaction, veracity, and aesthetic to make the heuristics more comprehensive, as the original 10 heuristics do not focus as much on these aspects of data visualization design [57].

Heuristic-guided expert review has some advantages as an evaluation method. On top of heuristics being able to be used by experts to preemptively design around these heuristics, the evaluation process is more structured and rapid compared to evaluation through controlled experimentation, which requires more time and resources. However, heuristic evaluation is not a replacement for user studies. Tory et al. (2005) and a follow-up study by Forsell (2012) recognized heuristic evaluation by experts may lead to identification of issues that are nonproblematic for general end users or may miss an issue that cannot be predicted without testing on a larger population of nonexperts [58,59].

The more holistic method for evaluating data visualization is to perform user studies, which directly approach issues that may arise with how an actual end user may use a data visualization. Wu et al. (2019) performed a systematic literature review of 76 publications that evaluated visual analytics for health informatics applications. Some studies used quantitative measures to evaluate attributes such as accuracy or efficiency of a visualization, while others collected subjective feedback to evaluate attributes that are harder to quantify such as user satisfaction. Many of the publications in their literature review, however, had an interest in collecting both quantitative and qualitative data to evaluate multiple aspects of a visualization. This was often done using task-based measurements such as time or accuracy to collect quantitative performance data in tandem with user feedback in the form of either open-ended responses or scoring on a scale [60]. One of the publications reviewed in this study tested the optimal method of graphical risk communication format to present data with small probabilities. This study used both quantitative and qualitative measures in the forms of timing task performance and user's rating of a visualization on a scale to compare whether user preference affected task performance [61]. In another example of evaluating health informatics information, Saraiya et al. (2005) compared five visualizations of gene expression based

on when a user drew their first insight and last insights from the visualization, how much participants felt they learned, and if participants felt that the visualizations answered all the questions that they had [62]. These measures are again both quantitative and qualitative in nature to explore both the objective effectiveness of the visualization as well as the user's opinion.

One example of user studies for evaluation of data visualization of physical activity data, Epstein (2016), compared seven different methods of visualizing step data from wearable fitness devices to determine which method would result in the lowest likelihood of a user lapsing [12]. Participants who had previously lapsed in tracking their activity used the seven visualizations to represent their own personal fitness data. These participants then gave their opinions on the design, elaborating on whether it was useful to them and other qualitative questions about their opinions on the design. The major difference between the evaluation in Epstein (2016) and this thesis is the lack of quantitative measurements to evaluate effectiveness. Epstein (2016) concentrated on evaluating user engagement rather than ability to draw insights, using perspective as the main criteria for good visualization design.

Although human subject testing is generally effective, great consideration must be put into the design of an experimental evaluation. Evaluations commonly struggle with issues such as unclear evaluation goals that make results difficult to interpret, pursuing effectiveness without defining it or considering all variables that may contribute to effectiveness or designing an evaluation with tasks that are not consistent with the goals of the evaluation [63]. Patterns in how data are collected can also be problematic with respect to the validity of data. Measuring only task times or errors can compromise validity in that a participant can perform a task quickly, but with an incorrect response. Post-task interviews may also miss the collection of important data if the participant cannot remember all the details of their experience performing that task. Although think

aloud protocol will help fill in these gaps in data, it may impact participant behavior [64]. North (2006) explains that task-based evaluation also does not reflect a visualization's true ability to offer insight, which is considered the core purpose of data visualization by many authors. Both the rigidity of predetermined tasks and the fact that these tasks end once the user finds the answer limit the possibility of unexpected insight [65].

Some general evaluation design approaches have been developed to guide the process of visualization evaluation. Lam et al. (2012) created seven guiding scenarios of visualization evaluation based on over 800 visualization publications. The second scenario described, evaluating visual data analysis and reasoning, pertains to investigating how a visualization tool can support the analytic process by observing both quantifiable metrics such as the number of insights gained during interactions with the tool and subjective feedback such as the user's satisfaction with the tool. The sample questions from this study that identify what should be considered during evaluation include: 1) how does the tool support seeking, searching, and extracting information? 2) how does the tool support hypothesis generation? And 3) how does the tool support decision making [66]? These sets of guiding questions as well as the designs of other experiments discussed in this section were referenced for the design of this thesis to establish how visualizations would be evaluated.

CHAPTER 3

METHODS

3.1 Overview

The methods by which the research questions posed by this study were answered were based on the heavily on the methods and findings of literature about evaluating data visualization design discussed in the previous section. I designed a mixed-methods human subjects testing experiment combining task-based subject response with post-task interviews to obtain quantitative and qualitative feedback. This study was approved by the University of Massachusetts IRB. This study approach directly addresses the following research questions:

1. How does the format of physical activity data visualizations impact older adults' abilities to infer meaning from physical activity data? Ability to interpret meaning is broken down into two key factors:
    a. Task success
    b. Task time
2. Are there other variations among users that impact an older adult's ability to infer accurate meaning from physical activity data? Variations considered include:
    a. Demographics (level of education, career field, gender)
    b. Current level of physical activity
    c. Level of experience with fitness tracker technology
    d. Graphical literacy

This study's design was also structured around the limitations of being conducted entirely remotely via Zoom. Remote sessions with participants were deemed the safest

option to eliminate participants' risk of exposure to COVID-19. Zoom was selected as the platform to conduct this remote research due to screen sharing features and the ability to record audio and screen sharing from a meeting, enabling collection of vast amounts of data with minimal inconvenience to the participant.

3.2 Participants

All participants recruited into the study were required to meet three inclusion criteria: 1) be aged 55 or older, 2) have a functioning computer that is connected to the internet with Zoom installed, and 3) feel proficient in the use of Zoom and internet browsers. The first criteria were established to target a population nearing or beyond age 65 since this population is soon entering or already part of a demographic with higher susceptibility to coronary heart disease [1,2]. This subset of the general population is an important stakeholder group for whom the potential benefits of mobile technologies encourage more active behavior are potentially significant, as exercise has been shown to positively impact cardiac health and decrease risk of developing heart disease [3-5]. The other two criteria for participation in this study were established because this study was carried out during the COVID-19 pandemic and that all interactions during the session would need to be carried out over Zoom.

30 participants were recruited for this study. This number of participants was selected based on it this thesis being a pilot study to get initial evaluation of which different data visualization design options tend to be most successful. Sessions lasted approximately two hours and participants were compensated with a $40 Amazon eGift card that was distributed through email.

3.3 Equipment & Materials

Several tools and technologies were required to carry out this experiment due to its remote nature. The items can be broken up into two major categories: physical equipment and virtual content.

3.3.1 Physical Equipment

The only physical equipment required to carry out this experiment were two computers with minimal accessories required. The roles and features of each piece of equipment are listed below:

- Computer 1 (participant's computer): must have working internet, mouse & keyboard, and microphone in addition to an installation of Zoom and a web browser
- Computer 2 (researcher's computer): must have working internet, mouse & keyboard, and microphone in addition to an installation of Zoom and a web browser; must also be able to host and share the visualization mockups and links to questionnaires and spreadsheets.

3.3.2 Virtual Content

Throughout the study session, the researcher and participant both interacted with several virtual tools and files. These tools and files included:

- Visualization mockups: three different physical activity data visualization mockups, discussed in greater detail below.

- Surveys: five different surveys meant to collect a variety of supplemental information; these surveys can be found in Appendices A through E and are listed below
    - Task Difficulty Questionnaire
    - Simple Physical Activity Questionnaire (SIMPAQ)
    - Physical Activity Tracking Experience Questionnaire
    - Demographics Questionnaire
    - Short Graphical Literacy Scale (GLS)

The Simple Physical Activity Questionnaire (SIMPAQ) and Short Graphical Literacy Scale (GLS) used in this study were selected as tools due to their validation in measuring physical activity levels and graphical literacy of individuals [50,52], and the demographics questionnaire follows a standard format that lists baseline demographic characteristics for education level, career field, gender, and an open field for health conditions. The task difficulty questionnaire and physical activity tracking experience survey were both created specifically for this study to collect supplemental. The task difficulty questionnaire was designed to allow participants to assign a relative difficulty rating to each task with the addition of only two other questions to clarify which visualization these ratings applied to, and which group this participant was a part of. The physical activity tracking experience questionnaire was designed to collect data on if the participants has engaged in fitness tracking experience at all, for how long, and with which apps and devices to account for any potential influence from experience with fitness tracking.

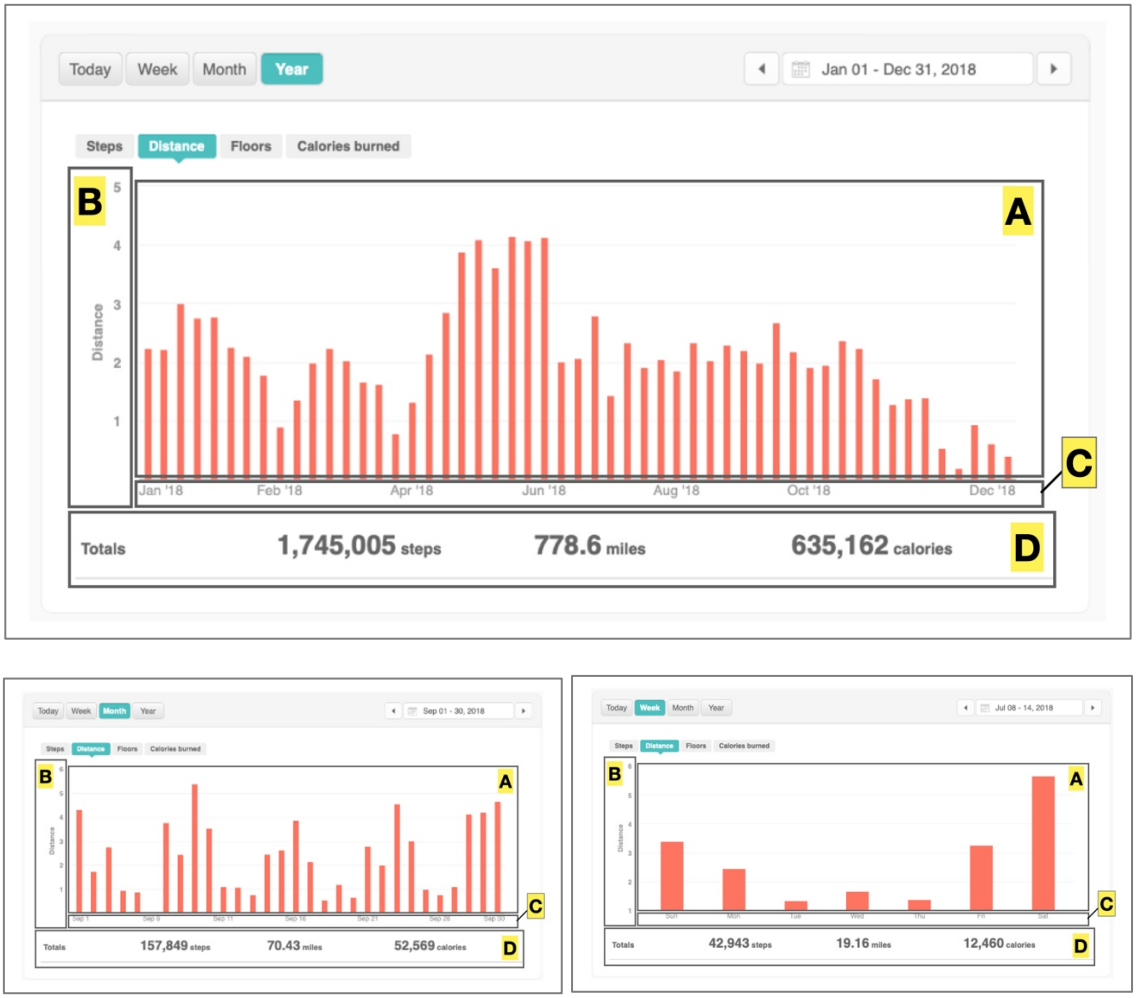
The visualization mockups used in this experiment included a mockup of Fitbit's desktop data visualization from their website, Strava's training log, and a prototype developed in Tableau called the Exercise Calendar. The data visualizations from Fitbit

and Strava were selected because they are two of the most widely used apps that focus on tracking physical activity [67], and the design of these two data visualizations does not overlap allowing evaluation of more design options. Screenshots of these data visualizations and details of how these data visualizations differ can be seen in Figure 1, Figure 2, and Figure 3. The mockups for Fitbit and the prototype were both created in Microsoft PowerPoint since they are navigated using button controls. A slide was created for each screen with interactive buttons mimicked as hyperlinks to different slides. The Strava training log mockup was captured in the form of a pdf file since this visualization has no button controls for navigation and is presented as a single continuously scrolling visualization on Strava's website. All interactions and functionality not pertaining to navigating through data chronologically were removed from mockup to prevent participants from navigating away from the data visualizations and to prevent participants from obtaining additional information not presented directly on the data visualizations.

Fitbit's physical activity data visualization design is centered around a horizontal bar graph that visually encodes distance walked per day or average distance walked per day over a week as bar length. The visualization is presented at three timespans: year-long, month-long, and week-long. Figure 1 shows screenshots of each timespan level of the Fitbit data visualization and breaks them down into four separate regions. Region A contains the bars that represent distance walked. Region B presents the y-axis that specifies the magnitude of distance for each bar. Region C contains the x-axis that specifies the date associated with each bar. Region D shows the timespan summary which includes totals over the entire timespan for steps, distance, and calories burned.

Figure 1

Breakdown of components in Fitbit data visualization



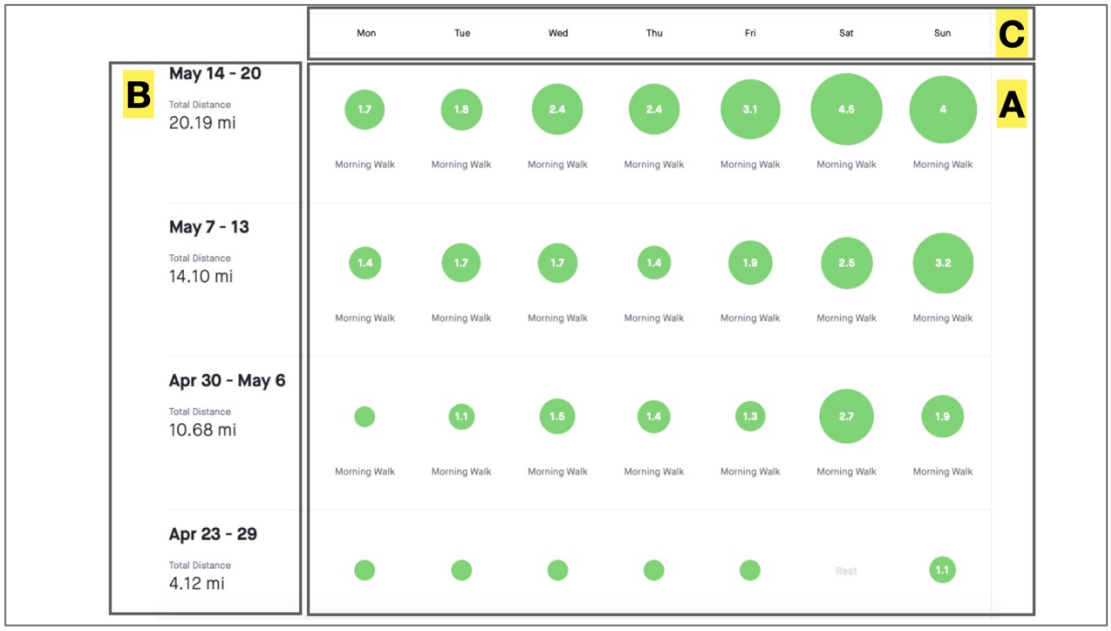
Note: The graph on the top is Fitbit's graph for a one-year timespan, the bottom left graph is for a month-long timespan, and the bottom right is for a one-week timespan. All three timespans share the same lettered components.

Strava's training log visually encodes distance walked per day as circle diameter and labels the distance value for each day with a distance greater than one mile on the circle. Days are arranged vertically by week, with the most recent week at the top and the least recent at the bottom, and days are arranged horizontally by day of the week, with Monday on the far left and Sunday on the far right. The visualization is only

presented at one level of granularity, with the ability to continuously scroll through all weeks in order. Figure 2 shows a screenshot of the Strava training log and breaks it down into three separate regions. Region A contains circles and labels that represent the total miles walked in a day. Region B presents summary information for the date range of the circles to the right as well as the total distance traveled over the course of that week. Region C is the headers for the days of the week of the circles in region A.

Figure 2

Breakdown of components in Strava data visualization



The prototype visualization is designed with two key methods of visual encoding at two different timescales. For the year overview, bar length is used to visually encode the distance traveled over the course of an entire month. For the month overview, bar length is used to visually encode the distance traveled over the course of a week, and color saturation is used to visually encode the distance walked on a day. Figure 3 shows screenshots of each timespan level of the prototype data visualization and breaks them

down into seven separate regions. Region A contains the colored squares arranged as a calendar that indicate miles walked on the corresponding date. Region B is the color key that provides reference for the distance associated with the color of squares in region A. Region C is the headers for the days of the week of the squares in region A. Region D presents horizontal bars and labels for the total miles traveled in the week for the dates of the squares in region A. Region E contains the bars and labels for the total miles walked in a month in the prototype's year overview. Region F is the y axis for the bars in Region E. Region G is the x axis that indicates which month the bars in region E correspond to.

Each of these three data visualizations display data from one year, and the data are displayed as the distance in miles the user walked during that period. The data used to create these data visualizations were generated by entering a year's worth of data into a spreadsheet and creating instances where certain values or patterns would occur at specific desired times. Each data visualization was built from different data sets with the same overall patters to ensure that participants would not begin memorizing the exact answers or general location of data points after their first and second set of task questions. The data points of interest that were involved in task questions were also controlled for difficulty. For example, questions that asked for comparing between data points maintain a similar relative difference in values between visualizations, and questions asking for a reading of a value keep that value within 10% of other values. More details on how data were controlled can be found in Appendix F.

Figure 3

Breakdown of components in prototype data visualization

Note: The graph on the top is the prototype's month overview and the bottom graph is for the prototype's year overview. These two timespans are visualized differently, and each graph's components are broken down separately.

3.4 Subject Testing Procedures

Beginning with recruitment of participants, interested participants were able to contact the study coordinator after they received information about the study from either a direct email to a listserv of individuals likely meeting the inclusion criteria, a Facebook post, an

organization called Rare Patient Voice, [68] or word-of-mouth from prior participants (snowball sampling). Potential participants were provided with a full outline of their participation in the study through email and were then invited to ask any questions or suggest a time to hold a session if they were interested.

Study participants were randomly assigned into one of three groups based on a Latin square design. These groups determined the order in which participants viewed and interacted with each of the data visualization mockups and ensured that each visualization was viewed first, second, or third in equal proportions (see Table 1). This approach was necessary to mitigate the risk of participants becoming more efficient with answering task questions as the session continued. Assuming there was some effect from the order of presentation, this experimental design approach prevents one visualization from artificially performing better because it is always the final visualizations mockup to be used.

Table 1

Visualization mockup order by group number

Group	First Mockup	Second Mockup	Third Mockup
Group 1	Fitbit	Strava	Prototype
Group 2	Strava	Prototype	Fitbit
Group 3	Prototype	Fitbit	Strava

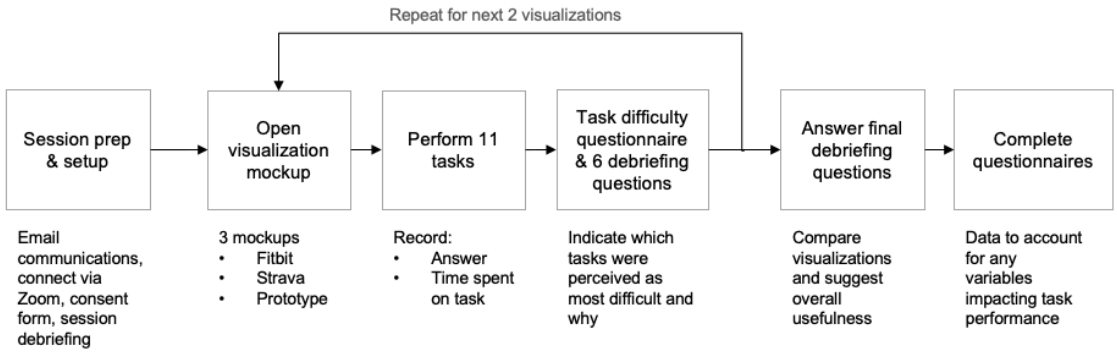
3.4.1 Preparation & Setup

Figure 4 provides an outline of all steps that occur following recruitment of a participant. Once a participant agreed to take part in the study and a time is chosen for them to hold their session, a Zoom call invitation link and a link to the consent form were sent via

email. A copy of this session confirmation email can be found in Appendix G. Participants were invited to read and sign the consent form prior to the study or wait until the session time and sign it at the beginning of the session. Participants were also assigned their random group and participant ID code at the time the session confirmation email was sent.

Figure 4

Flowchart of participant session procedures



At the time of the session, the researcher started the Zoom meeting room and waited for the participant to enter the room. Once a participant read, understood, and signed the consent form, they were provided an overview of the general structure of the session as well as how to approach task questions. Participants were also informed that audio and video would begin recording at that time, after which the recording would begin. The notes that the researcher used to debrief the participant at the start of the Zoom meeting can be found in Appendix H. The first visualization mockup was then opened from the researcher's computer and screen sharing with remote control was enabled for the participant over Zoom so the participant could view and interact directly with the first

visualization mockup through the Zoom call window without needing to open any other applications.

3.4.2 Visualization Tasks and Debriefing Interview

Once a visualization was available to the participant to interact with through Zoom, the participant began completing the task questions. Because physical activity data visualizations focus on questions an individual may have about their own exercise habits or history, a wide variety of task question types were considered. To guide what questions individuals may have about their own data and which questions should therefore be included in the study, Brehmer and Muzner's task typology for visualization tasks [69] was used as a foundation to consider a wide variety of manners in which a user may seek insight into their activity behaviors. Brehmer and Muzner's search types were unaltered, but the task types were slightly adjusted to suit the nature of this study and the query type "interpret" was added. Interpret as a query type implies a necessity to accurately decode the visual encoding used in a specified data point. Table 2 breaks down questions by task types. The process for completing task questions was as follows:

1. Researcher verbally asked one task question (see Table 2) and started a stopwatch.
2. Participant explored the data visualization to find an answer.
3. Participant verbally responded with an answer to the task question.
4. Researcher ended stopwatch and recorded their response and the stopwatch time in their data entry worksheet.

Table 2

Breakdown of task types for all task questions

Task	Question	Search Type	Query Type	Resolution
1	During which month was Jane most active?	Browse	Compare	Month (sum)
2	During which month was Jane least active?	Browse	Compare	Month (sum)
3	In June, on which two days of the week did Jane tend to be most active?	Browse	Compare	Days (pattern)
4	In June, on which three days of the week did Jane tend to be least active?	Browse	Compare	Days (pattern)
5	What was the date of Jane's most active day in July?	Browse	Compare	Day (value)
6	What was the date of Jane's least active day in July?	Browse	Compare	Day (value)
7	How far did Jane walk in the entire month of February? You may round to the nearest whole mile.	Lookup	Interpret	Month (sum)
8	How far did Jane walk during the week of November 5th? You may round to the nearest whole mile.	Lookup	Interpret	Week (sum)
9	How far did Jane walk on May 12th? You may round to the nearest whole mile.	Lookup	Interpret	Day (value)
10	In September, on how many days did Jane walk less than 1 mile?	Browse	Interpret	Days (value)
11	At some point in 2018, Jane missed collecting data for 9 days in a row. What is the start date of this 9-day streak?	Locate	Identify	Days (value)

Steps 1 – 4 were repeated until all task questions were completed. Immediately following completion of all task questions, participants were asked to complete a brief task difficulty questionnaire that was available online through Qualtrics. A copy of the contents of this survey can be found in Appendix A. The purpose of this questionnaire was to assign a Likert scale value to indicate how difficult each task question felt for the participant to complete. This task difficulty questionnaire was then followed by six interview questions that were intended to understand the results obtained during task completion as well as gain a general impression of what participants did and did not like about the visualization. The six interview questions were as follows:

1. Which task question (or questions) was easiest for you to answer? Why?
2. Which task question (or questions) was most difficult for you to answer? Why?
3. Does this visualization have any strengths in your opinion? These can be, but are not required to be, related to the task questions that you answered.
4. Does this visualization have any weaknesses in your opinion? These can be, but are not required to be, related to the task questions that you answered.
5. Are there any things that you would remove from this visualization?
6. Are there any things that you would add to this visualization?
7. Are there any other things that you would change about this visualization?

These questions were asked verbally by the researcher and answered verbally by the participant. The researcher took notes to summarize the participant's response in addition to the audio recording being collected by Zoom. Participants' responses to these questions were not used as analytical data, only obtained in this study to describe possible performance explanations and may be coded into themes as a separate study to gain further insights. Once the interview questions for one data visualization were completed, participants were presented with the next visualization to interact with until all three visualizations were complete.

3.4.3 Final Interview

Once tasks and debriefing interviews for all three visualizations were completed, the researcher conducted a final debriefing interview to understand what physical activity data were more or less important to participants. The questions asked in this interview were as follows:

1. Which visualization do you like the best? Why?
2. Which visualization do you like the least? Why?

3. Do you currently use any personal health and fitness tracking devices or track health and fitness manually?
    a. What devices if any, and what do you record (ex: steps, calories burned, sleep)?
    b. If you do not track but had personal health and fitness data available to you (assume no additional effort), what would you track?
4. What are (or would be) your goals relating to personal health and fitness tracking? Are these goals related to any health conditions?
5. Do you have an interest in viewing past data collected through health and fitness devices?
6. Do you think any of the data visualizations that we looked at would be helpful in accomplishing the goals you outlined?
7. What health and fitness tracking motives do you think the data visualizations that you interacted with today fail to facilitate?
8. Do any of the task questions relate to questions you would want to answer with your own health and fitness data? What questions that you would ask are not represented in the task questions?

As with the debriefing interview questions, participants' responses to these questions were not used as analytical data, only obtained in this study to describe possible performance explanations and may be coded into themes as a separate study to gain further insights into topics such as which of the data visualizations provided each participant with the overall user experience that they found most and least desirable along with supporting reasons for that preference, and how personal tracking devices and data factor into each participant's lifestyle and motivations. Like the debriefing interviews, these questions were asked verbally by the researcher and answered

verbally by the participant. The researcher took notes to summarize the participant's response in addition to the audio recording being collected by Zoom. Participants were encouraged to expand as much as they like on any of the questions to fully explain their thoughts and opinions.

3.4.4 Questionnaires

Several factors may influence how individuals use the visualizations, so participants completed four brief surveys using Qualtrics to collect data for these factors. These surveys were the simple physical activity questionnaire (SIMPAQ), fitness tracking experience survey, demographics questionnaire, and short graphical literacy scale (GLS). These questionnaires can all be found in Appendices B through E. Participants were told they could skip any questions that they did not feel comfortable answering. After they were done filling out the questionnaires, the audio recording on Zoom was terminated, the participant was thanked for their time, and a $40 Amazon gift card was emailed to the participant.

3.5 Analysis Approach

All quantitative data collected throughout this study were analyzed through summarization and statistical tests to answer the research questions posed by this study. Table 3 is a list of the variables were measured, including information about what type of variable it was, how it was collected, what the potential values were, and their general method of analysis. At the end of this section, table 4 provides a summary of all analyses conducted on the data collected for this study.

Table 3

List of variables and descriptions of how they were collected and used

Measure Name	Measure Type	Measurement Method	Potential Values	Analysis used
Visualization mockup	Independent (categorical)	Recorded as note	Fitbit, Strava training log, Tableau prototype	N/A
Visualization interaction position	Moderating (categorical)	Random assigned, recorded as note	First, second, third	N/A
Task success	Dependent (categorical)	If participant response matches correct answer	Successful, unsuccessful	Pairwise chi square test, logistic regression
Task time	Dependent (continuous)	Timing from when task is given until response is given by participant	From zero seconds upwards	Tukey-Kramer test between visualizations parsed by task
Task difficulty rating	Dependent (ordinal)	Questionnaire responses (custom questionnaire for this study)	Very easy, somewhat easy, somewhat difficult, very difficult	Logistic regression by task time
Interview responses	Dependent (categorical)	Audio recording of participant responses to interview questions, note taking	N/A	Coding for themes, tallying themes
Prior physical activity tracking experience	Moderating (continuous)	Questionnaire responses (custom questionnaire for this study)	Years spent tracking	Correlations with task performance
Graphical Literacy	Moderating (ordinal categorical)	Questionnaire responses (GLS)	0 to 4 correct responses	Correlations with task performance
Physical activity level	Moderating (continuous)	Questionnaire responses (SIMPAQ)	Avg hours of active time per typical day, from 0 to 24 hours	Correlations with task performance
Demographic information • Career field • Education level • Gender • Ethnicity • Chronic health conditions	Moderating (categorical)	Questionnaire responses (demographic questionnaire)	N/A	Correlations with task performance

The first data to be reviewed was task data: correctness of participant responses, task completion times, and task difficulty ratings. Participant responses were reviewed, and incorrect responses were tallied by task and visualization. These totals for incorrect

responses were compared across visualizations for each task and examined for a statistically significant difference through pairwise chi square tests between visualizations. Visualizations that had statistically significant differences in incorrect responses were identified and categorized as performing better or worse for that task. The answers themselves were analyzed next to identify patterns in incorrect responses. The number of unique incorrect answers were tallied by task and by visualization, and then the unique responses that occur 3 times or more were identified and designated as a common mistake.

Task times were summarized by mean task time and standard deviation for each combination of task and visualization. These values were used to perform Tukey HSD test between visualizations for each task to identify visualizations with a statistically significant difference in task time distributions. These visualizations were categorized in terms of better or worse task time performance in the same manner as how the visualizations were categorized for task response correctness. The relationship between mean task time and incorrect task responses was also explored through linear regression to understand if time to complete a task was a reliable predictor of whether the task was completed successfully.

Task difficulty ratings from participant questionnaire responses had each response tallied between visualizations for each task. Difficulty ratings were quantified as factor levels and examined for a relationship with incorrect task responses and mean task time for each task/visualization combination through linear regression.

Following the analyses outlined above, the second research question posed by this study examined the variety of potentially confounding variables that were considered in this study. Beginning with demographic information including education level, career field, sex, and ethnicity, the distribution of individuals falling into each category in each of the randomly assigned groups was analyzed through a chi square test to check for any

imbalance between the composition of each group. For the other categorical variables of whether a participant has previously engaged in physical activity tracking or what types of activity tracking devices they have used, the same form of analysis was used. Continuous variables such as a participant's average amount of hours of physical activity in a day, the amount of time they have spent tracking physical activity in the past, or the number of questions they correctly answered on the Graphical Literacy Scale, an ANOVA test was used.

Once control variables were tested for evenness of distribution across the three randomly assigned interaction order groups, they were tested for influence on task time and successful task completion. For all control variables, they were tested individually for task and visualization combinations as well as across the entire overall dataset. To test the influence of categorical control variables over correct task response, which includes the categorical variables listed above plus assigned interaction order group, chi square tests were used. To test the categorical control variable influence over task time, ANOVA tests were used. To test the continuous control variable influence over correct task responses and mean task time grouped by task and visualization, linear regression was used.

While the debriefing interview questions were not analyzed using systematic qualitative approaches, participant responses were used to try to understand how visualization design led participants to their response that led to resulting quantitative measurements. Future work will include an in-depth analysis to of these important qualitative data.

Table 4

List of analyses performed and statistical tests used

Analysis Step	Independent Variable	Dependent Variable	By Variables	Test Used
1	Visualization	Proportion of incorrect task responses	Task	Chi square test
2	Visualization	# unique incorrect task responses	Task	N/A
3	Visualization	Task time	Task	Tukey HSD
4	Mean task Time	Proportion of incorrect task responses	Visualization + task	Correlation
5	Visualization	Task difficulty rating	Task	N/A
6	Proportion of incorrect task responses	Mean task difficulty rating	Visualization + task	Correlation
7	Mean task time	Mean task difficulty rating	Visualization + task	Correlation
8	Group	Proportion of incorrect task responses	Visualization, task (separately)	Chi square test
9	Group	Mean task time	Visualization, task (separately)	ANOVA
10	Group	Proportion of group belonging to different demographic groups	None	Chi square test
11	Group	Mean values of continuous control variables	None	ANOVA
12	Demographic group	Proportion of incorrect task responses	Visualization, task (separately)	Chi square test
13	Demographic group	Mean task time	Visualization, task (separately)	ANOVA
14	Continuous control variable	Proportion of incorrect task responses	Visualization + task	Correlation
15	Continuous control variable	Mean task time	Visualization + task	Correlation

CHAPTER 4

RESULTS

Following the analysis approach outlined in section 3.5, this results section will separate analyses by the research question they are intended to answer. Restating the two research questions that this study seeks to answer:

1. How does the format of physical activity data visualizations impact older adults' abilities to infer meaning from physical activity data? Ability to interpret meaning is broken down into two key factors:
    a. Task success
    b. Task time
2. Are there other variations among users impact an older adult's ability to infer accurate meaning from physical activity data? Variations considered include:
    a. Demographic characteristic (level of education, career field, gender)
    b. Current level of physical activity
    c. Level of experience with fitness tracker technology
    d. Graphical literacy

4.1 Task Performance

The primary means of addressing how the format of physical activity data visualizations impact older adults' abilities to infer meaning from physical activity data is to compare the performance of each of the three visualizations. Figures 5 and 6 below give an overview of the relative performance of each data visualization's performance for the number of incorrect task responses and mean task time for each task.

Figure 5

Number of incorrect task responses for each visualization across all 11 tasks

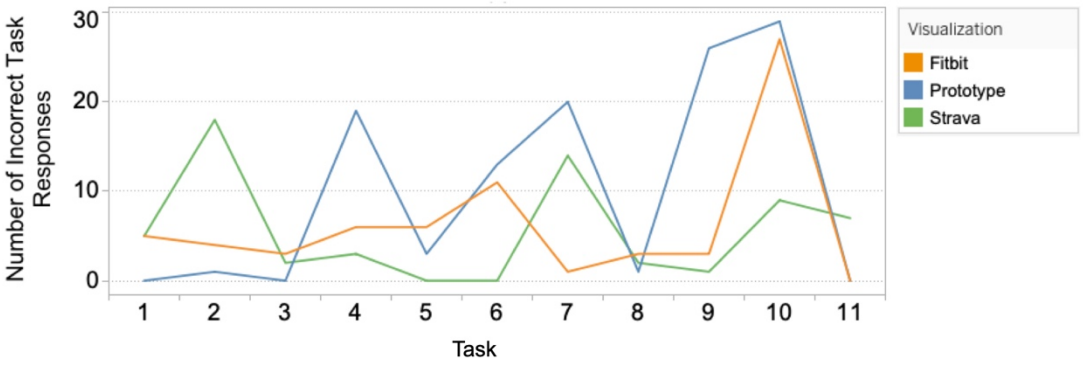


Figure 6

Average task time for each visualization across all 11 tasks

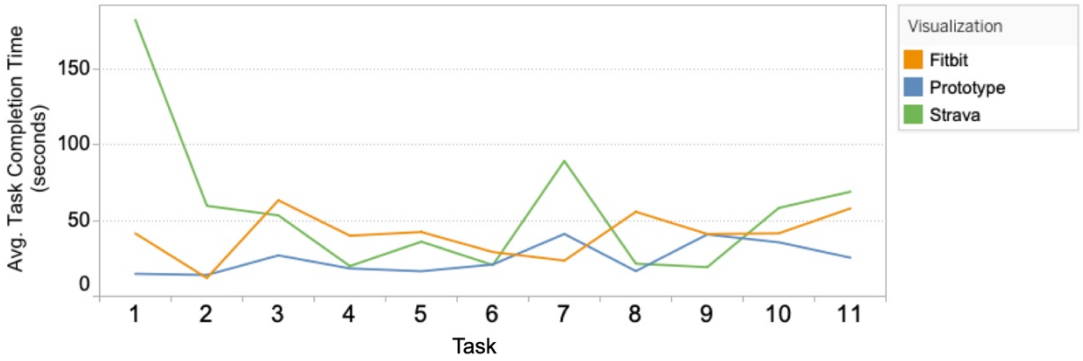


Figure 5 illustrates several key observations. First is that all three visualizations share low incorrect response rates for task 3 and task 8. Chi square analysis performed for these tasks included in Table I1 of Appendix I supports this finding, showing that for these tasks all visualizations had a statistically similar proportion of incorrect responses. Figure 5 also shows that none of the three visualizations were able to maintain a low incorrect response rate for all 11 tasks, with the prototype most frequently having the

highest number of incorrect responses on tasks. Strava most often had the fewest incorrect responses across tasks and was the least frequent to have the highest number of incorrect responses.

In Figure 6, the task time distributions were similar across all three visualizations for task 6 only. This is supported by Table J1 in Appendix J in which Tukey HSD tests were conducted for the task time distributions of each visualization on each task. Figure 6 also shows the prototype's mean task times were most frequently the lowest of the three visualizations for the 11 tasks. The prototype also never had the highest mean task time on any of the 11 tasks. The two highest mean task times belonged to Strava, but Fitbit's mean task times were most often the highest.

For remaining task and metric combinations that had a more variable performance, performance for each visualization will be presented separately. Three tables are provided with a breakdown of the task performance of each visualization for each task and indicate when a visualization performs significantly better or worse on a task. Table 5 summarizes the task performance of the Fitbit visualization, Table 6 summarizes the task performance of the Strava visualization, and Table 7 summarizes the task performance of the prototype visualization. Within these tables, the results of particular interest are incorrect response counts of five of more, incorrect response counts that are statistically higher or lower than other visualizations for the same task, mean task times that are roughly a minute or longer, and task time distributions that are statistically higher or lower than other visualizations for the same task. The performance of visualizations in these instances will be explored in the discussion section of this paper.

Table 5

Summary of Fitbit's task performance for all 11 tasks

Task	Question	Search Type	Query Type	Resolution	Incorrect Responses	Task Time	
						$\bar{x}$	s
1	During which month was Jane most active?	Browse	Compare	Month (sum)	**5**	41.5	39.8
2	During which month was Jane least active?	Browse	Compare	Month (sum)	4	12.1	11.8
3	In June, on which two days of the week did Jane tend to be most active?	Browse	Compare	Days (pattern)	3	**63.3**	**40.1**
4	In June, on which three days of the week did Jane tend to be least active?	Browse	Compare	Days (pattern)	**6**	**40.0****	**39.1**
5	What was the date of Jane's most active day in July?	Browse	Compare	Day (value)	**6**	42.4	42.4
6	What was the date of Jane's least active day in July?	Browse	Compare	Day (value)	**11**	29.2	28.5
7	How far did Jane walk in the entire month of February? You may round to the nearest whole mile.	Lookup	Interpret	Month (sum)	**1***	23.6	24.8
8	How far did Jane walk during the week of November 5th? You may round to the nearest whole mile.	Lookup	Interpret	Week (sum)	3	**55.8****	**94.9**
9	How far did Jane walk on May 12th? You may round to the nearest whole mile.	Lookup	Interpret	Day (value)	3	41.1	27.2
10	In September, on how many days did Jane walk less than 1 mile?	Browse	Interpret	Days (value)	**27**	41.5	27.4
11	At some point in 2018, Jane missed collecting data for 9 days in a row. What is the start date of this 9-day streak?	Locate	Identify	Days (value)	0	**58.0**	**46.6**

* Visualization performed better for metric than other visualizations on given task

** Visualization performed worse for metric than other visualizations on given task

Note: Better/worse performance on task for incorrect response rate determined by chi square tests in Appendix I; better/worse performance for time determined by Tukey HSD test in Appendix J. Bold text without asterisk denotes task performance that does not have a statistically significant difference from other visualizations but is still of interest.

Table 6

Summary of Strava's task performance for all 11 tasks

Task	Question	Search Type	Query Type	Resolution	Incorrect Responses	Task Time	
						$\bar{x}$	s
1	During which month was Jane most active?	Browse	Compare	Month (sum)	**5**	**182****	**171**
2	During which month was Jane least active?	Browse	Compare	Month (sum)	**18****	**59.7****	**46.7**
3	In June, on which two days of the week did Jane tend to be most active?	Browse	Compare	Days (pattern)	2	53.4	57.2
4	In June, on which three days of the week did Jane tend to be least active?	Browse	Compare	Days (pattern)	3	19.9	10.6
5	What was the date of Jane's most active day in July?	Browse	Compare	Day (value)	**0***	36.0	23.1
6	What was the date of Jane's least active day in July?	Browse	Compare	Day (value)	**0***	20.6	10.8
7	How far did Jane walk in the entire month of February? You may round to the nearest whole mile.	Lookup	Interpret	Month (sum)	**14**	**89.3****	**47.3**
8	How far did Jane walk during the week of November 5th? You may round to the nearest whole mile.	Lookup	Interpret	Week (sum)	2	21.5	11.1
9	How far did Jane walk on May 12th? You may round to the nearest whole mile.	Lookup	Interpret	Day (value)	1	**19.3***	**13.3**
10	In September, on how many days did Jane walk less than 1 mile?	Browse	Interpret	Days (value)	**9***	**58.3****	**25.1**
11	At some point in 2018, Jane missed collecting data for 9 days in a row. What is the start date of this 9-day streak?	Locate	Identify	Days (value)	**7****	**69.0**	**72.9**

* Visualization performed better for metric than other visualizations on given task

** Visualization performed worse for metric than other visualizations on given task

Note: Better/worse performance on task for incorrect response rate determined by chi square tests in Appendix I; better/worse performance for time determined by Tukey HSD test in Appendix J. Bold text without asterisk denotes task performance that does not have a statistically significant difference from other visualizations but is still of interest.

Table 7

Summary of Prototype's task performance for all 11 tasks

Task	Question	Search Type	Query Type	Resolution	Incorrect Responses	Task Time	
						$\bar{x}$	s
1	During which month was Jane most active?	Browse	Compare	Month (sum)	**0**	**14.8**	**57.2**
2	During which month was Jane least active?	Browse	Compare	Month (sum)	**1**	**14.1**	**59.5**
3	In June, on which two days of the week did Jane tend to be most active?	Browse	Compare	Days (pattern)	0	**27.0***	**22.8**
4	In June, on which three days of the week did Jane tend to be least active?	Browse	Compare	Days (pattern)	**19****	18.4	18.0
5	What was the date of Jane's most active day in July?	Browse	Compare	Day (value)	3	**16.5***	**15.3**
6	What was the date of Jane's least active day in July?	Browse	Compare	Day (value)	**13**	21.0	22.0
7	How far did Jane walk in the entire month of February? You may round to the nearest whole mile.	Lookup	Interpret	Month (sum)	**20**	41.2	36.5
8	How far did Jane walk during the week of November 5th? You may round to the nearest whole mile.	Lookup	Interpret	Week (sum)	1	16.6	11.6
9	How far did Jane walk on May 12th? You may round to the nearest whole mile.	Lookup	Interpret	Day (value)	**26****	40.9	22.2
10	In September, on how many days did Jane walk less than 1 mile?	Browse	Interpret	Days (value)	**29**	35.6	22.9
11	At some point in 2018, Jane missed collecting data for 9 days in a row. What is the start date of this 9-day streak?	Locate	Identify	Days (value)	0	**25.5***	**21.5**

* Visualization performed better for metric than other visualizations on given task

** Visualization performed worse for metric than other visualizations on given task

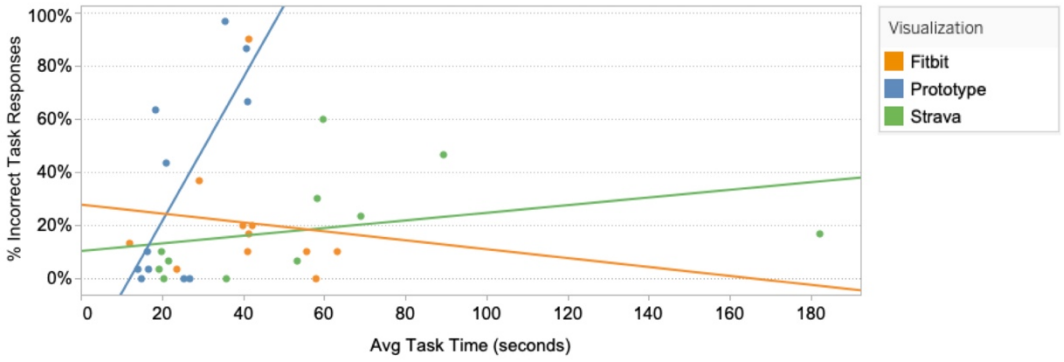
Note: Better/worse performance on task for incorrect response rate determined by chi square tests in Appendix I; better/worse performance for time determined by Tukey HSD test in Appendix J. Bold text without asterisk denotes task performance that does not have a statistically significant difference from other visualizations but is still of interest.

4.1.1 Time vs Likelihood of Incorrectness

The impact of task time on the likelihood that a participant incorrectly answered a task question was also examined. Figure 7 plots the average task time for each task and visualization against the percent of incorrect task responses for that task (in this case referred to as % task error). Using these data points, a correlation test with a linear trend line was fitted to the data points for each visualization separately to establish if a relationship between the two variables existed. The results show a weak correlation of these variables for Fitbit and Strava ($p = 0.765$ for Fitbit, $p = 0.297$ for Strava), but that the prototype had a strong correlation for these variables ($p = 0.011$). The correlation is also positive, meaning that participants took more time to answer questions that they were more likely to have an incorrect response for.

Figure 7

Average task time vs % incorrect task responses for each visualization across all 11 tasks with trend lines, correlation p values, and correlation formulas

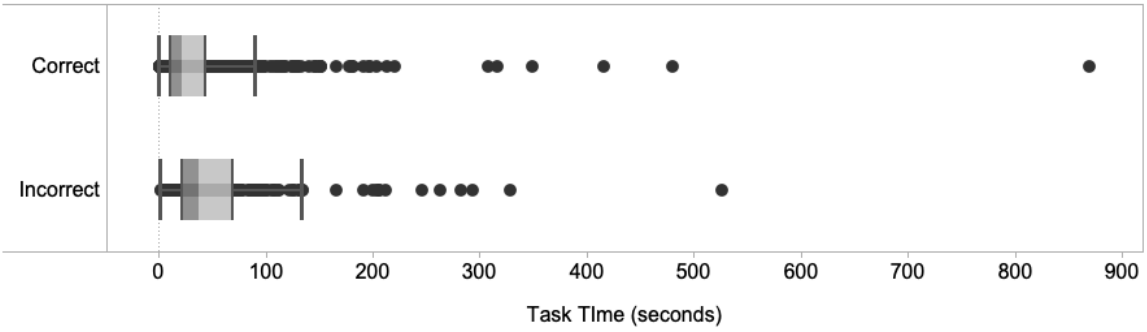


Visualization	Correlation p value	Formula
Fitbit	7.65E-01	Incorrect Task Response Rate = -0.0016824*Avg. Time + 0.277677
Strava	2.97E-01	Incorrect Task Response Rate = 0.00144274*Avg. Time + 0.102317
Prototype	**1.13E-02**	Incorrect Task Response Rate = 0.0271024*Avg. Time + -0.329748

Figure 8 below shows a box and whisker plot for the task times associated with each completed task and separated by if that task response was correct or incorrect. Looking at Figure 8, it is apparent that the distribution of mean task times is similar regardless of whether the task responses are correct or incorrect. Performing the same analysis as in Figure 7 without separating results by visualization yields a weak correlation ($p = 0.568$).

Figure 8

Distributions of task time for correct or incorrect task response



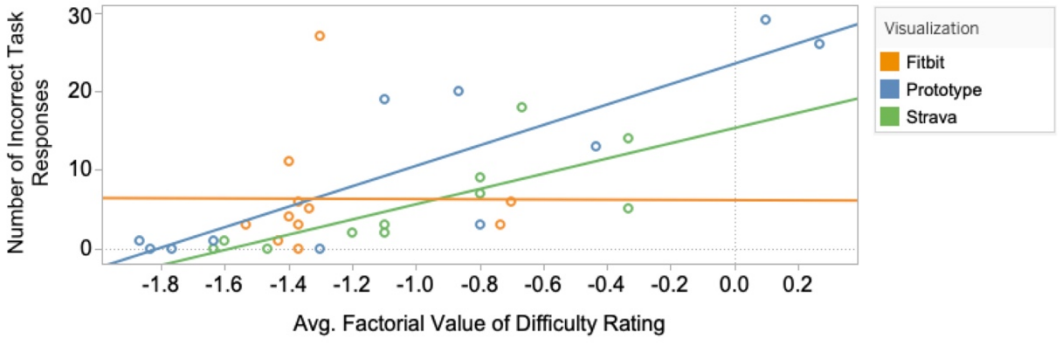
4.1.2 Task Difficulty

The difficulty ratings assigned by participants to each task question for each visualization were tallied and consolidated into a table in appendix K. The tallies for each visualization show that tasks using Fitbit were generally considered easier than the other two visualizations. Strava and the prototype both had more instances in which tasks were considered more difficult. However, tasks completed with the prototype were more generally assigned an extreme difficult rating such as “Very Easy” or “Very Difficult” while Strava was more frequently assigned moderate difficulty rating of “Somewhat Easy” or “Somewhat Difficult.”

Figure 9 shows a scatterplot of incorrect task responses against average task difficulty factor level for each task, separated by visualization. A factor level was assigned to each difficultly level rating to allow the ordinal variable of difficulty rating to be treated as a continuous variable and analyzed quantitatively. “Very Easy” tasks were given a factor value of -2, “Somewhat Easy” was given a value of -1, “Somewhat Difficult” was given a value of 1, and “Very Difficult” was given a value of 2. Factors were arranged this way to frame higher ratings as more difficult so that a positive correlation with incorrect response rate would mean higher incorrect response rates occur with higher difficulty. The results of these regressions in Figure 9 indicate that tasks that had higher rates of incorrect answers also tended to be rated as more difficult for tasks completed using Strava or the prototype.

Figure 9

Scatterplot of incorrect task responses against average task difficulty factor level for each task, separated by visualization

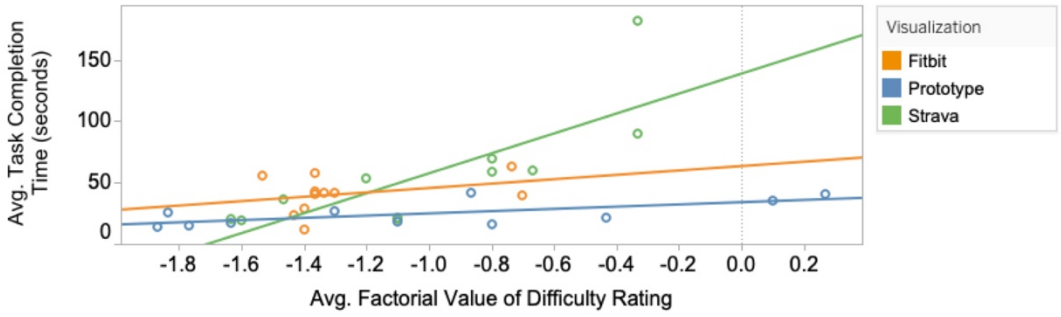


Visualization	Correlation p value	Formula
Fitbit	9.89E-01	Incorrect Task Responses = -0.12894*Avg. Difficulty Factorial + 6.1094
Strava	**7.90E-03**	Incorrect Task Responses = 9.69966*Avg. Difficulty Factorial + 15.2745
Prototype	**7.44E-04**	Incorrect Task Responses = 13.008*Avg. Difficulty Factorial + 23.4657

Figure 10 shows a scatterplot of average task completion time against average task difficulty factor level for each task, separated by visualization under the same methodology as in Figure 9. The results of these regressions indicate that tasks that had longer average task times also tended to be rated as more difficult for tasks completed using Strava or the prototype.

Figure 10

Scatterplot of average task completion time against average task difficulty factor level for each task, separated by visualization



Visualization	Correlation p value	Formula
Fitbit	3.24E-01	Avg. Task Time = 17.8611*Avg. Difficulty Factorial + 63.391
Strava	**4.27E-03**	Avg. Task Time = 81.1898*Avg. Difficulty Factorial + 138.64
Prototype	**2.37E-02**	Avg. Task Time = 9.10254*Avg. Difficulty Factorial + 33.985

4.2 Impacts of Other Factors

The second research question that this study asks is how traits of older adults or other factors influence their ability to read and interpret physical activity data visualization. The results of the relationships between these traits and factors and task performance with each visualization were analyzed separately and are presented in this section.

4.2.1 Research Group

Appendix L provides a series of tables that examine any statistically significant differences in task performance between groups. Breaking down the number of incorrect task responses between each group for each task and visualization and comparing these numbers through chi square tests revealed that group had in impact on the number of incorrect responses overall when using the Fitbit data visualization ($p = 0.044$). Group 1, the group that uses Fitbit as the first visualization, had significantly more incorrect responses than group 2 and group 3. However, comparing the task time distributions between groups by task and by visualization using ANOVA tests yielded no statistically significant differences for any of the distributions.

4.2.2 Demographics

Task performance differences between different baseline demographic characteristics were compared in the same manner as for research groups: number of incorrect task responses between each baseline demographic characteristic for a given task or visualization were compared through chi square tests, and task time distributions between each baseline demographic characteristic for a given task or visualization were compared through ANOVA tests. The summary data and results of these analyses are contained in appendix M. Incorrect response numbers when separated by task did not result in any statistically significant differences between any of the baseline characteristics. However, when separated by visualization, the number of incorrect task responses had a statistically significant difference based on career field for the Fitbit visualization ($p = 0.015$). Analyzing differences in task time distributions between baseline characteristics by task revealed a statistically significant difference between

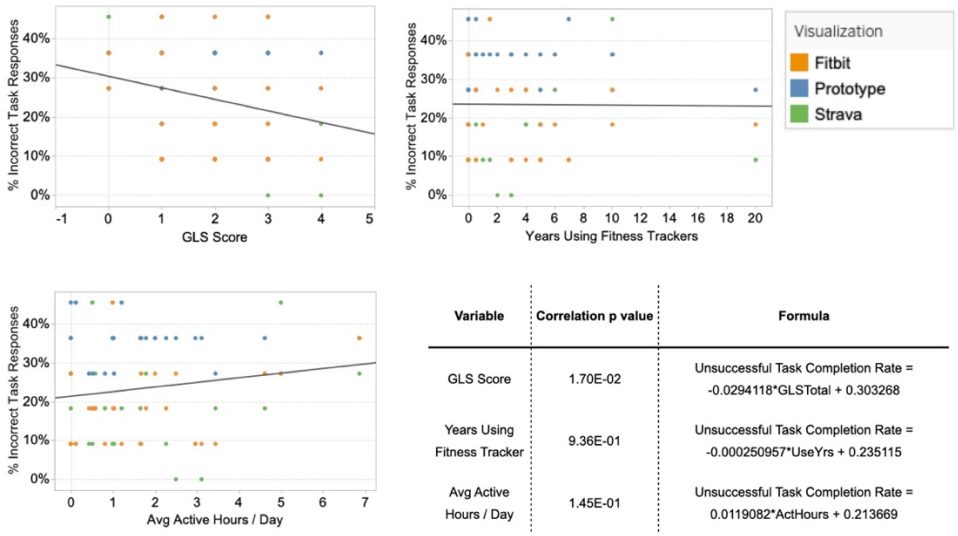
baseline groupings for education on task 8 ($p = 0.027$), for career field on task 9 ($p = 5.7e-05$), and for gender on task 9 ($p = 0.043$). Analyzing differences in task time distributions between baseline characteristics by visualization also revealed a statistically significant difference between baseline groupings for education when using Strava ($p = 0.011$) and when combining all task and visualization data together ($p = 0.05$). It is important to note that many of the samples for baseline demographic characteristics were very small, and so this analysis may not be reflective of differences in performance based off these demographic characteristics within the full population of older adults.

4.2.3 Other Factors

Figure 11 below shows scatterplots of the percent of incorrect task responses for each visualization against the continuous moderating variables: GLS score, year using a fitness tracker, and average active hours per day. Years using a fitness tracker and average active hours per day show a weak correlation with number of incorrect task responses ($p = 0.94$ and $p = 0.15$). However, GLS score shows a strong correlation with the number of incorrect task responses ($p = 0.017$) suggesting that for every question on the GLS a participant answered correctly, they were roughly 3% more likely to have an incorrect response to a task question.

Figure 11

Scatterplots with trend lines for GLS Score (top left), Years Using Fitness Trackers (top right), and Avg Active Hours / Day (bottom left) vs incorrect task response rate

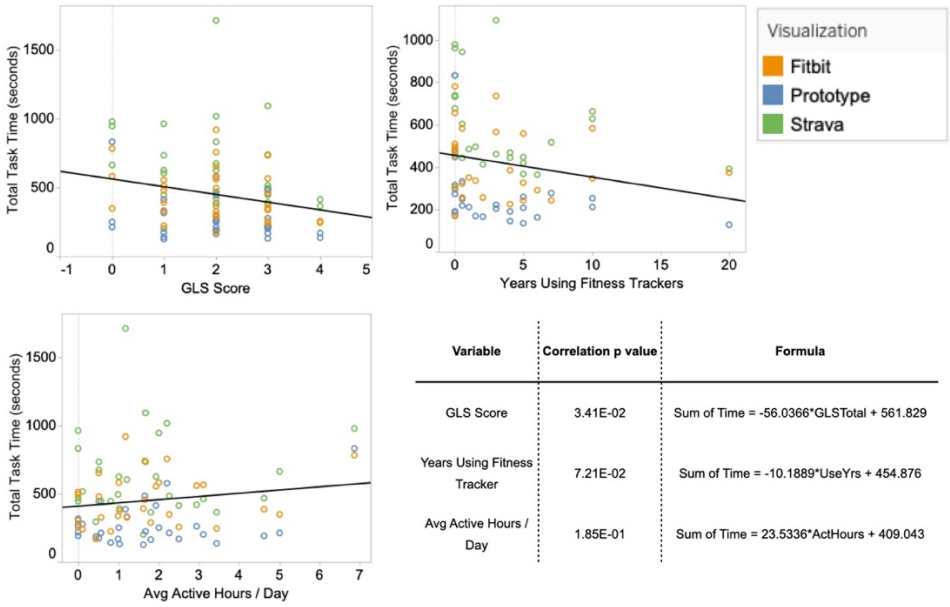


Variable	Correlation p value	Formula
GLS Score	1.70E-02	Unsuccessful Task Completion Rate = -0.0294118*GLSTotal + 0.303268
Years Using Fitness Tracker	9.36E-01	Unsuccessful Task Completion Rate = -0.000250957*UseYrs + 0.235115
Avg Active Hours / Day	1.45E-01	Unsuccessful Task Completion Rate = 0.0119082*ActHours + 0.213669

Figure 12 below shows scatterplots of total task completion times for each visualization against the continuous moderating variables: GLS score, year using a fitness tracker, and average active hours per day. Years using a fitness tracker and average active hours per day show a weak correlation with mean task time ($p = 0.072$ and $p = 0.16$). However, GLS score shows a strong correlation with mean task time ($p = 0.034$) suggesting that for every question on the GLS a participant answered correctly, they were likely to have taken 56 seconds less to complete all 11 tasks for a visualization.

Figure 12

Scatterplots with trend lines for GLS Score (top left), Years Using Fitness Trackers (top right), and Avg Active Hours / Day (bottom left) vs task completion times



Variable	Correlation p value	Formula
GLS Score	3.41E-02	Sum of Time = -56.0366*GLSTotal + 561.829
Years Using Fitness Tracker	7.21E-02	Sum of Time = -10.1889*UseYrs + 454.876
Avg Active Hours / Day	1.85E-01	Sum of Time = 23.5336*ActHours + 409.043

4.2.4 Group Composition

The demographic makeup of each of the three visualization order groups was checked to monitor for any statistically significant differences in makeup that may have influence over group task performance. Appendix N provides tables for the chi square test results of the distribution categorical demographic variables between groups and ANOVA results for the distribution of continuous factors between groups. All tests resulted in statistically insignificant p values aside from the ANOVA for GLS distribution (p = 1.57e-04). GLS distribution showed group 2 having a higher GLS score distribution and group 1 having the lowest.

CHAPTER 5

DISCUSSION

The outcomes of task performance in chapter 4 provided quantitative results that roughly outline the limitations of what information each visualization can provide to older adults. These findings are valuable in answering the question of how the format of physical activity data visualizations impact older adults' abilities to infer accurate meaning from physical activity data. However, the intent of this study to gain a thorough understanding of how data visualization formats impact older adult's ability to infer meaning from those visualizations also necessitates understanding what interactions users have with that data visualization and how those interactions result in a correct or incorrect response, and how much effort is required by the user to arrive at that response. Participant responses and feedback during task questions and interviews throughout the study are the starting point for gaining this understanding. As mentioned in Chapter 3, the debriefing interview questions were not analyzed using systematic qualitative approaches, but participant responses were used to try to understand how visualization design led participants to their response that led to resulting quantitative measurements.

5.1 Explanations of Task Performances

As previously described in Section 3.3, Fitbit's fundamental data visualization format is a time series bar graph in which each bar represents the miles walked on a single day or average miles walked in a week. The different basic components of this data visualization, which were shown back in Figure 1, include the bars that represent distance walked, the y-axis that specifies the magnitude of distance for each bar, the x-

axis that specifies the date associated with each bar, and the timespan summary which includes totals over the entire timespan for steps, distance, and calories burned. These components each played a different role as participants explored data to complete task questions, and were used differently for different tasks, sometimes having a positive influence on task completion and other times having a negative influence.

This discussion of task performance will begin with the x axis, which is intended to present information about the date or date range that each bar corresponds to. When participants searched for answers to task questions 1 and 2 using Fitbit's visualization, many participants struggled to use the x axis for the year overview to identify during which month in the year the most and least active weeks occurred. The placement of data labels does not follow a regular interval, with some adjacent labels being for consecutive months and others being a full month apart. These labels also were not consistently aligned with bars representing a particular point in that month, and no other visual cue to signify the start or end of a month was present. When narrowing the scope to the month overview, the x axis presented labels for the dates of some of the bars but lacked labeling to signify what day of the week a bar corresponded with. The week overview's x axis was the opposite, displaying labels on the x axis for the day of the week a bar represented but not the date. Fitbit's results for tasks 3 and 4, in which participants were asked to establish a week-to-week pattern in activity, suggested that participants needed more time to view the separate week overviews to establish which days of the week regularly had the most or least activity since the month overview did not provide all the information needed to answer this task question.

Some design choices for Fitbit's y axis also had negative impacts on overall task performance. The range of the y axis varies depending on the range of miles across the dates that are being displayed. For example, in a week during which the fewest miles walked on a single day was just under two miles, the y axis would begin at one mile

rather than zero miles. Good Charts by Scott Berinato explains that truncating a y axis in this manner carries the risk of exaggerating the difference in values between different bars [70]. This can be advantageous to emphasize differences in bars, but the visual representation of actual values for each bar will not be accurate in addition to the visual representation of the differences between these values. The inaccuracy of this visual representation of differences is compounded in situations where users compare the length of bars on different screens where the y axis has different ranges of values. The length of a bar representing 1.5 miles with a y axis starting a one mile may appear shorter than a bar representing 0.9 miles on a y axis starting at zero miles depending on the maximum value of each y axis, meaning that users would need to use the y axis to estimate the approximate value of each bar and then compare these values. Literature about graphical literacy suggests that users having lower graphical literacy may have trouble approximating these values or even may not notice that change in y axis ranges and make comparisons of miles walked based on bar length alone [46]. This exact scenario occurred multiple times during task 6 when participants interacted with the Fitbit visualization, during which eight of the thirty participants believed a date with 1.3 miles walked had less walking distance than a date with 0.6 miles because the week of the date with 1.3 miles had a y axis beginning at one mile and the date with 0.6 miles was on a week where the y axis began at zero miles.

Of the 11 tasks that participants were asked to complete, task 10, in which participants needed to determine the number of days in September on which less than one mile was walked, was Fitbit's worst task performance in terms of the number of incorrect responses. 27 participants had an incorrect response to this task question, and 23 of those participants responses were incorrect at least partially because participants could not determine that 0.99 miles were walked on the date of September 25th, instead reading the bar to be equal to one mile. The difficulty of reading the exact value of this

bar may have been made even more difficult by the low contrast ratio of the horizontal reference line to the background of the bar graph. Individuals aged 60 or older have been shown to have a reduced contrast sensitivity [54]. The horizontal reference lines marking each whole mile on the y axis are a light gray color having a contrast ratio of only 1.09:1 on the white background of this bar graph. Many participants remarked either that they did not notice these reference lines for several task questions or remarked that they would have liked the addition of these reference lines not realizing that they were already present at all. Though the importance of knowing a value to the exact hundredth of a mile may not exist for all users, these results still signify that using the bars alone in Fitbit's visualization cannot reliably present data to this level of detail.

One of the most successful aspects of Fitbit's data visualization design was the summaries provided under the bar graph that present totals for the steps and miles walked in the period being visualized as well as the estimated calories burned. Many participants used these summaries to complete tasks 7 and 8 which ask for total miles walked during the month of February and the week of November 5th, respectively. Since these totals are written as a number rather than being visualized in some other manner, participants could simply navigate to the screen for the correct time frame and reference the number of miles at the bottom of the screen.

Strava's visualization method was very different from Fitbit's, using circle size instead of bar length to visualize miles walked on an individual date and arranging dates of the same week in the same row and dates of the same day of the week vertically in the same column with labels for the miles walked on an individual day to the nearest tenth of a mile. This visualization was comprised of three major components, shown back in Figure 2, which are the circles and labels that represent the total miles walked in a day, the headers for the days of the week of the circles, and the summary information for the date range of the circles to the right as well as the total distance traveled over the

course of that week. The range of dates and sum of miles for the week, which were positioned to the left of the corresponding week were the primary reason that Strava's visualization had a strong performance for task 8 which asked participants to determine the total number of miles walked during the week of November 5th. Strava's performance on tasks requiring assessment of the miles walked on a particular date such as task 9 or a comparison of the miles walked on multiple dates such as tasks 3 through 6 also generally had very few incorrect responses due to the fact that participants could quickly identify the dates with the most or least activity based on circle size and then compare the values of the labels between circles with similar sizes or read the label on the circle corresponding to the date of interest. However, on dates where less than one miles was walked, data labels for miles walked are not present. Task 10, though similar in nature to task 9, had more incorrect responses than task 9 because participants were unsure about the meaning of small circles without labels and how they differed from dates labeled "Rest."

Participants had difficulty completing Task 11 in which they were asked to find a nine-day streak of no recorded activity for the same reason as difficulty in completing task 10, being the ambiguous meaning of the small circles with no label, in addition to the fact that the nine-day streak did not appear to be continuous because some participants read dates in the incorrect order. Dates are displayed with the most recent weeks at the top of the page and less recent weeks appearing as the user scrolls down, and days of the week are arranged in order from left to right. This means that if a user wants to view their activity from three consecutive weeks in order, they need to scroll down to the first date they want to view, read from left to right until the rightmost day, and then move up a week and read from left to right again. Several participants viewing dates out of order is not surprising when considering that a Gregorian calendar is read from top to bottom and left to right rather than bottom to top and left to right.

Another task type that participants struggled to complete when using Strava's visualization were tasks concerning the total activity in a month because of Strava's lack of summarization of miles walked at this time span. To respond to tasks 1 and 2 and identify the most and least active months, participants needed to either judge the amount of activity visually by typical circle sizes or take more time and estimate the miles walked during that month using the weekly totals to the left. Both techniques were shown to be prone to error and to be time consuming. The same applied to task 7 which asks for the total miles walked in the month of February. Participants could arrive at this value be adding the weekly totals for the weeks that were entirely contained within February and then adding the individual dates from February on weeks where part of the week was in January or March. However, many participants performed this calculation mentally and had an incorrect response due to errors in rounding or arithmetic.

The final of the three visualizations, the prototype, uses two different visualization methods to display data: a simple bar graph for the total miles walked in each month for a year overview, and a calendar format that colors each date on the calendar a darker shade of green the more activity occurs on that date. Each of these two visualization methods is comprised of several components, which are shown back in Figure 3. To restate, the components of the heatmap visualization include the colored squares arranged as a Gregorian calendar that indicate miles walked on the corresponding date, the color key that provides reference for the distance associated with the color of squares in the calendar, the headers for the days of the week of the squares of the calendar, and horizontal bars and labels for the total miles traveled in the week for the dates shown on the calendar directly to the left. In the year overview, components included are the bars and labels for the total miles walked in a month, the y axis for the bars, and the x axis that indicates which month the bars correspond to.

Tasks 1 and 2 were quickly and correctly responded to because participants were able to quickly compare bar lengths to see which months had the most or least activity and could reference the labels on each bar that show the total miles for that month. Participants were also able to quickly and correctly obtain the miles walked during the week of November 5th for task 8 since bars with labels for total miles walked during that week are displayed to the right of the calendar week in the month overviews. Despite the prototype performing well on tasks 1 and 2, this visualization did not perform as well on task 7 in which participants were asked to figure out the total miles walked in the month of February. Although this total was available as a data label for the bar representing February in the month overview, 20 out of participants preferred to use February's month overview page. Participants added up the week totals shown in that month since no monthly summary values are provided in the month overview pages, but since the week totals include miles from days that are at the end of the previous month or beginning of the next month, this strategy did not result in a correct response from the participants who used it.

The prototype's performance on tasks that are concerned with assessing the miles walked on a day or a comparison of miles walked on different days was variable, particularly for the number of incorrect responses. Because the miles walked on a day are represented by the darkness and saturation of the color in the square for that date, to compare how many miles were walked on different days requires a user to compare the colors in each of the squares and determine which is lighter or darker. As mentioned when discussing Fitbit's horizontal reference lines, older adults tend to have lesser contrast sensitivity [54], so making comparisons between similar colors in tasks 4 and 6 was difficult where the contrast ratios were 1.06:1 and 1.01:1 respectively when comparing the most similar values that were relevant to the task questions. The comparison for task 3 was less difficult than for task 4 due to a higher contrast ratio of

1.35:1 for the second most active day of the week and third most active day of the week. Although task 5 had a contrast ratio of 1.06:1 like task 4 did, it had far fewer incorrect responses. Appendix O includes a table listing the colors being compared in these tasks and their contrast ratios for reference. It is possible that because the comparison was only between two days in that month which were both colored significantly darker than all surrounding days that it was easier to compare these colors, but the reason for performing better on this task when comparing the same contrast ratio is unclear at this time. Tasks 9 and 10, which require assessment of the color of squares on the calendar to determine the miles walked on those days had the highest incorrect response counts of all tasks when completed using the prototype. The only reference that can be used to associate color with miles walked is a key to the left of the calendar that shows a color gradient with the minimum and maximum value labeled on either end of the gradient. In general, participants were not able to estimate within a mile of the actual value of miles walked on dates of interest.

5.2 Task Difficulty vs Task Time & Task Response

The analysis section compared mean task time and the percentage of incorrect task responses for each visualization-task pair to the average difficulty rating for that pairing. That analysis revealed that for tasks with higher percentages of incorrect task responses and higher average task completion times, participants rated those tasks as more difficult on average for Strava and the prototype. This finding suggests that when using these visualizations to complete certain tasks, participants were generally aware when they spent a longer time on a task or when they did not answer the task correctly and may also suggest (although not definitively) that participants are aware of the reasons or

design elements that cause task completion to take more time or have a higher likelihood of incorrect task response.

5.3 Other Impactors of Task Performance

Of the array of potentially moderating variables considered by this study, some were found to have influence over task outcomes. There were several instances in which variability in the sample showed statistical significance, but the only moderating variable that was found to impact both task response correctness and task completion time across all tasks and visualizations was GLS score. Graphical literacy is expected to have an inherent affect the outcomes of task response correctness and task completion time due to it being a direct measure of how well an individual can read a data visualization [45], and that tasks in this thesis are centered around viewing and interpreting data visualizations. High graphical literacy implies that an individual can properly read a data visualization and that they are able to do so more accurately and efficiently than someone with lower graphical literacy [46]. However, in general, a larger sample size is needed to make a concrete determination of the effects of the variables among the intended population that are highlighted in this study. When comparing performance between education level or career, for example, some of these subgroups were as small as two or one person. No strong evidence was found supporting any variable aside from graphical literacy having a strong impact on older adults' ability to interpret these different forms of data visualizations, but this thesis also cannot put forward the absence of such an influence.

5.4 Design Recommendations

The differences and similarities in results of each visualization across task questions and the probable reasons for these results offers substantial insight into what works well and what does not for visualizing an individual's physical activity history for a user base of older adults. The explanations of task performance point out many design choices that have a positive impact in answering various questions while others have a negative impact, but three principles of design stand out as the most universal and beneficial:

1. Make exact values available
2. Summarize data at multiple timescales
3. Ensure accessibility for the entire population

These three principles of visualization design for physical activity data have some overlap, but each has its own significance for designing a successful data visualization of physical activity data for older adults.

5.4.1 Make Exact Values Available

The most consistent result across all visualizations was tasks whose correct answer was written as a number had many fewer incorrect responses. Providing numbers for users to reference eliminates the need for a user to translate a visual feature into a quantified value, allowing a user to answer questions about their data more efficiently and accurately. However, this does not mean that written numbers should replace visually encoding values. Strava's performance on tasks 3 through 6 showed that combining visual encoding with labels for exact numbers can be a powerful tool that facilitates accurate and efficient review of data. When completing these tasks with Strava, participants were able to quickly scan the size of the circles to find dates that they were interested in, whether large or small, and use the labels to compare the exact values

between dates when circle sizes were too similar to compare alone. Though in Fitbit and Strava's native forms, this information is already available by interacting with the dates of interest on each visualization, the results obtained by removing these features from the visualization validate how important the presence of this function is to a successful data visualization of this physical activity data.

5.4.2 Summarize Data at Multiple Timescales

The only data visualization used in this study that did not have at least two different timescales available to view data was Strava training log. When participants completed the tasks that required comparison of the amount of activity in different months or obtaining the total number of miles walked in a month, this visualization's limited ability to help a user answer these types of questions became apparent with high numbers of incorrect responses and long task completion times. Having no visualization of data totaled over longer periods of time, such as total distance walked over a month, limits the ability of users to compare the amount of activity across longer periods of time, see long-term trends, and presence or absence of long-term patterns that may repeat from year to year.

5.4.3 Ensure Accessibility for the Entire Population

Considering the universal benefit of regular physical activity on individuals' health and the potential of physical activity tracking technology to improve physical activity habits [3-5], data visualizations for physical activity tracking should be designed in such a way that all users can utilize these visualizations to their full potential and gain all possible benefits. Users of such technology may have a variety of characteristics that limit or

negate the usefulness of the visualization due to accessibility issues. Two such characteristics that were already mentioned in this study are low graphical literacy, in which an individual has a lesser ability to read and interpret data visualizations, and low color contrast sensitivity, in which individuals are less capable of distinguishing between similar colors. Another relevant characteristic is colorblindness, of which there are many different forms that limit an individual's ability to distinguish between certain color pairings, or in some cases, detect colors altogether. Have awareness of how these characteristics impact accessibility and putting care info format and color palette choices for data visualization designs can ensure that the visualization is equally accessible to all individuals. The United States General Services Administration (GSA) Technology Transformation Services has developed standards regarding minimum acceptable contrast ratio of text on webpages which can be applied to text and other features on a visualization's design [71], and tools exist to simulate how an image appears to individuals with different forms of color blindness so the color palette can be inspected for overlap in perceived color across forms of color blindness [53].

5.5. Prototype Improvements

Though the prototype was designed specifically for this thesis, there were still several tasks with poor performance results. The three major design recommendations from above can be applied to the prototype to identify design solutions that will strengthen the weak points of this visualization and make it a more effective tool overall. Beginning with making exact values available, a means of obtaining the exact number of miles walked on a day should be provided for users. One option is to record the exact number of miles on each square along with the color coding used to signal general level of activity, but this may clutter the visualization. Instead, making the visualization dynamic and allowing users to get details about what amount of physical activity occurred on a date by either

hovering a cursor over it or clicking on it will provide users with all the information they need to know how much walking occurred on an individual day and will facilitate accurate comparison of the amount of walking that occurred among several different dates.

Now considering summarizing data at multiple timescales, the prototype already provides visualizations that offer totals for miles walked at the resolution levels of months, weeks, and individual days. However, a total for the number of miles walked in a month on that month screen will negate the need for users to change to the year view if they want to know the total miles walked in a month while on the month overview, and it would also prevent the users from adding the week totals and arriving at an incorrect value for total miles walked in that month.

Lastly, to make the prototype visualization as accessible as possible, the addition of exact mile values in a popup for individual days is recommended again. Providing this information allows reduces a user's need to rely on graphical literacy or their sensitivity to color contrast. Also testing the current design and color palette for different types of color blindness, the colors for the bars representing total distance walked in a week and the colors for shading the individual dates on the calendar appears to overlap in hue for the color blindness types of tritanopia (blue-yellow color blindness) and monochromacy (total color blindness). With the current design, users could mistakenly add some significance to the shading of the bars for total distance in a week since the shading for the dates in the calendar portion has significance. The most effective resolution for this issue would be to remove the horizontal bar altogether and leave only the label for miles walked during that week similar to Strava's training log format. The length of the bar is somewhat redundant when acknowledging that the shading of individual dates in a week already fulfills the role of visually indicating that a week had significantly more or less activity than the preceding or following week. Users would presumably still be able to

scan and identify when a week had a relatively high or low amount of physical activity and reference the label for the total miles walked in a week to determine the differences between weeks that had similar levels of activity.

5.6 Recommendations for New and Existing Physical Activity Tracker Users

Although the focus of this thesis is on how to better design data visualizations of physical activity data to more readable and useful for older adults, this information can also be used to guide what apps and tools an older adult should adopt based on how physical activity data is visualized. As a user, first considering what goals you have for tracking physical activity will help you decide what app or tool is right for you. From these goals, come up with several questions that you may have about your own data like the 11 task questions used in this study. After either entering placeholder data into the application or collecting physical activity data, you can attempt to use the data visualizations generated in the application or tool to answer the questions that you came up with. If the data visualization seemed to allow you to answer all these questions easily and if the answers to these questions were informative or useful, the application or tool may be a good fit for collecting your physical activity data. Otherwise, a different tool may exist that is a better fit.

5.7 Incorporation of Additional Data

Many medical professionals and consumers have expressed interest in combining physical activity data with other personal health data to achieve a more holistic overview of one's health. Medical professionals have considered incorporating this data into EHRs to monitor patients' exercise habits since these habits may help to prevent or manage conditions such as heart disease and diabetes [23-25]. Participants from this study also expressed interest in combining physical activity data with additional data such as nutrition, weight, blood pressure, and blood glucose level to understand how exercise

habits correlate with or impact other measurements. These types of questions require more data and are more complicated to answer. Consequently, answering these questions through data visualization would require a different form of data visualization that can effectively visualize more than one type of measure. The key design recommendations for physical activity data visualization from this thesis are not intended to guide design of data visualizations of this nature and are limited to visualization only of measures that relate to physical activity history.

5.8 Limitations

This study has several limitations that are important to consider when incorporating the findings of this study to a more general application. The population of this study was not diverse: 100% of participants were white, 87% were female, and 93% have some form of college degree. Though this study was prepared to examine the effect of these different demographics on how individuals may interact with these data visualizations, how the findings of this study apply to the general population cannot be fully understood with this absence of variation in participant demographics. Unfortunately, data for more precise ranges in age of participants was not collected due to that field on the demographic questionnaire being lost in the migration of all questionnaires to Qualtrics for remote Zoom sessions. For the analysis of how the amount of fitness tracking experience participants have affects task performance, this may also not be accurate due to eight participants indicating that they had used fitness trackers but not inputting the amount of time for which they had used them.

The task performance data presented in this study also cannot be assumed to be reflective of the long-term utility or usability of these data visualizations. Participants in this study were not educated on how to read any of the data visualizations, and

questions regarding how to read them were not allowed to be answered during the session. For an actual application or data visualization used in the real world, users would have access to countless resources to help them learn to better read and understand the data that they are collecting. These resources could come in the form of an in-app tutorial, online forums, online video guides, or customer service to name a few.

Participants' task difficulty ratings also may have limited validity because of the time at which the questions to assign them were introduced. Participants were asked to assign task difficulty ratings for all 11 tasks after the final task was completed, and several participants verbalized that they could not recall how quickly or easily they arrived at an answer. Participants may have assumed better or worse performance on a task than what they had experienced.

The visualization mockups in this study did not include some features that may have been helpful in task completion for the sake of evaluating the visualizations as directly as possible. Visualizations housed in modern apps or websites could allow the user to hover over or click or tap on a data point on the data visualization to get an exact readout of the distance or other metric of interest that was achieved on that data point. Visualization mockups and tasks also only presented physical activity in the form of distance walked in miles. Many users may prefer a different metric such as steps walked, calories burned, or even distance represented in a different unit such as kilometers. Displaying data with these different metrics may have produced different results, and users may find these different metrics more useful or less useful than distance walked.

Lastly, this study makes no claims regarding the potential of any of these visualizations to better inspire older adults to perform the recommended amount of physical activity on a regular basis. This research paper only recognizes that improved

data visualization has the potential to inspire a greater level of physical activity by teaching users more about their habits and health from the data that they are collecting.

5.9 Future Work

Considering the limitations of this study, some opportunities for related future work exist. A longitudinal trial with one or more high-fidelity data visualizations developed based on the findings of this study would provide more insight into what visualizations have the best performance over long term, what older adults are able to learn from the data visualizations they interact with, what information older adults are most interested in learning from their data, and if different, more informative data visualizations can motivate older adults to be more physically active.

CHAPTER 6

CONCLUSION

In this thesis, a task-based human subject study was performed with three different data visualizations to gain insight into how the format of physical activity data visualizations impact older adults' abilities to infer meaning from physical activity data. Also considered were additional effects from other traits or factors that vary between individuals such as demographic group, previous experience with fitness tracking technology, and graphical literacy. A prototype data visualization that was designed specifically for this study and two data visualizations from popular on-market apps for physical activity tracking were used by participants to complete a set of 11 tasks. Results from the time taken to complete these tasks and responses given suggest that how data visualizations are formatted heavily affect how older adults infer meaning. Each visualization had its own unique strengths and limitations, with none exhibiting strong performance across all tasks. From the successes and shortcomings of each visualization, three key design recommendations for the design of data visualizations for physical activity data were made: 1) make exact values available, 2) summarize data at multiple timescales, and 3) ensure accessibility for the entire population. Although a few statistical tests showed a statistically significant relationship between moderating variables and task performance, these results did not show a strong pattern across visualizations or tasks. The efforts made in this study have provided some direction for how physical activity data visualizations can be better designed for older adult users, but there is still much more to be learned to create a more informative system that can inspire a more active lifestyle.

APPENDIX A

TASK DIFFICULTY QUESTIONNAIRE

Start of Block: Default Question Block

Q1 Which visualization did you just interact with?

- Fitbit (bar graphs)
- Strava training log (circles)
- Exercise calendar (colored squares)

Q2 Was this the first, second, or third visualization you interacted with?

- First
- Second
- Third

Q3 Below, please select the difficulty level you feel best represents the difficulty of completing each task question with the visualization that you just interacted with. (difficulty selection table on next page)

	Very Easy (1)	Somewhat Easy (2)	Somewhat Difficult (3)	Very Difficult (4)
Task 1: During which month of 2018 was Jane most active? (1)	○	○	○	○
Task 2: During which month of 2018 was Jane least active? (2)	○	○	○	○
Task 3: In June 2018, on which two days of the week did Jane tend to walk furthest? (4)	○	○	○	○
Task 4: In June 2018, on which two days of the week did Jane tend to walk least? (5)	○	○	○	○
Task 5: During the month of July 2018, on which day did Jane record the most distance walked? (6)	○	○	○	○
Task 6: During the month of July 2018, on which day did Jane record the least distance walked? (7)	○	○	○	○
Task 7: How many miles did Jane walk in February 2018? Round to the nearest whole number in miles. (8)	○	○	○	○
Task 8: How many miles did Jane walk during the week of November 5th, 2018? Round to the nearest whole number in miles. (9)	○	○	○	○
Task 9: How many miles did Jane walk on May 12th, 2018? Round to the nearest whole number in miles. (10)	○	○	○	○
Task 10: In September of 2018, on how many days did Jane walk under 1 mile? (11)	○	○	○	○
Task 11: At some point in 2018, Jane had a streak of 9 consecutive days with no physical activity recorded (no distance walked). What is the start date of this 9-day streak? (12)	○	○	○	○

End of Block: Default Question Block

APPENDIX B

SIMPLE PHYSICAL ACTIVITY TRACKING QUESTIONNAIRE (SIMPAQ)

This questionnaire will ask you about what you have been doing over the past seven days, including time spent in bed, sitting, or lying down, walking, exercise, sport, and other activities.

1A	What time did you usually go to bed over the past seven days?		ex: 10:00pm, 1:00am

1B	What time did you usually get out of bed over the past seven days?	

Average hours in bed per night	**0**	
That leaves approximately	24	*hours a day for other activities.*

2A	Out of those remaining hours, how long did you spend sitting or lying own, such as when you are eating, reading, watching TV, or using electronic devices?		hours		minutes
2B	How much of this time is spent napping?		hours		minutes

Total	**0**	
That leaves approximately	24	*hours a day for other activities.*

3	Which days in the past seven days did you walk for exercise or recreation or to get to or from places? How many minutes did you usually spend walking on those days?

Monday	**Tuesday**	**Wednesday**	**Thursday**	**Friday**	**Saturday**	**Sunday**

Avg hours walking per day	

4	Now think about an activity that you do for exercise and sport, such as jogging, running, swimming, bike riding, going to the gym, yoga, etc. Which days in the past week did you do any of the, or similar activities, and how long did you spend on each activity on each day? Please list below.

Day	**Activity and intensity (0-10)**	**Number of sessions**	**Minutes spent on each activity**	**Total minutes exercised**
example	resistance training (5/10); tennis (9/10)	1; 1	15; 50	65
Monday				
Tuesday				
Wednesday				
Thursday				
Friday				
Saturday				
Sunday				
			Total	0

Avg hours sport/exercise per day	**0**

5	Now think about any other physical activities that you did as part of your work or activities you did while at home such as gardening or household chores. How many minutes did you spend on these activities on most days?
	minutes/days

Sum of times (should be roughly 24 hours)	

APPENDIX C

PHYSICAL ACTIVITY TRACKING EXPERIENCE QUESTIONNAIRE

Start of Block: Default Question Block

Q1 Have you ever used a fitness tracking device or fitness tracking app before?

- Yes (1)
- No (2)

Display This Question:

If Have you ever used a fitness tracking device or fitness tracking app before? = No

Q2 Is there a particular reason you have not used a fitness tracking device or fitness tracking app before? If so, please explain below.

_____

Display This Question:

If Have you ever used a fitness tracking device or fitness tracking app before? = No

Q3 What features could make fitness tracking useful for you?

_____

Display This Question:

If Have you ever used a fitness tracking device or fitness tracking app before? = Yes

Q5 Do you still use your fitness tracker?

- ○ Yes (1)
- ○ No (2)

Display This Question:

If Have you ever used a fitness tracking device or fitness tracking app before? = Yes

Q4 For how long have you used fitness tracking technology?

_____

Display This Question:

If Have you ever used a fitness tracking device or fitness tracking app before? = Yes

And Do you still use your fitness tracker? = Yes

Q6 What keeps you engaged in tracking fitness/exercise?

_____

Display This Question:

If Have you ever used a fitness tracking device or fitness tracking app before? = Yes

And Do you still use your fitness tracker? = No

Q7 Why did you stop using your fitness tracking technology?

_____

Display This Question:

If Have you ever used a fitness tracking device or fitness tracking app before? = Yes

Q9 What apps and devices did you use?

_____

Display This Question:

If Have you ever used a fitness tracking device or fitness tracking app before? = Yes

Q8 Which of the following most closely describes how you tracked your fitness?

- ○ Smartphone only (1)
- ○ Wearable device on wrist (2)
- ○ Wearable device on waist (3)
- ○ Other (4) _____

Display This Question:

If Have you ever used a fitness tracking device or fitness tracking app before? = Yes

Q10 What data were you interested in tracking? Select all that apply.

- ☐ Time spent exercising (1)
- ☐ Distance moved (2)
- ☐ Steps taken (3)
- ☐ Calories burned (4)
- ☐ Heart rate (5)
- ☐ Sleep (6)
- ☐ Other (7) _____

End of Block: Default Question Block

APPENDIX D

DEMOGRAPHICS QUESTIONNAIRE

Start of Block: Default Question Block

Q1 Highest level of education completed?

- ○ No schooling completed (1)
- ○ Elementary school to 8th grade (2)
- ○ Some high school, no diploma (3)
- ○ High school graduate (diploma or equivalent) (4)
- ○ Trade / technical / vocational training (5)
- ○ Some college credit (no degree) (6)
- ○ Associate's degree (7)
- ○ Bachelor's degree (8)
- ○ Master's degree (9)
- ○ Professional degree (10)
- ○ Doctorate degree (11)
- ○ Prefer not to say (12)

Q2 Career or field of study?

- ○ Arts and humanities (1)
- ○ Business (2)
- ○ Health and medicine (3)
- ○ Public and social services (4)
- ○ Science, math, and technology (5)
- ○ Service (6)
- ○ Social sciences (7)
- ○ Trades and professional services (8)
- ○ Other (9) _____
- ○ Prefer not to say (10)

Q4 Do you identify as any of the below genders?

- ○ Male (1)
- ○ Female (2)
- ○ Non-binary / third gender (3)
- ○ Agender (4)
- ○ Prefer not to say (5)
- ○ Prefer to self-describe (6) _____

Q5 What ethnicity do you most closely identify with?

- Asian / Pacific Islander (1)
- Black or African American (2)
- Hispanic or Latino (3)
- Native American or American Indian (4)
- White (5)
- Other (6) _____
- Prefer not to say (7)

Q6 Have you been diagnosed with any chronic health conditions? If so, please list below.

_____

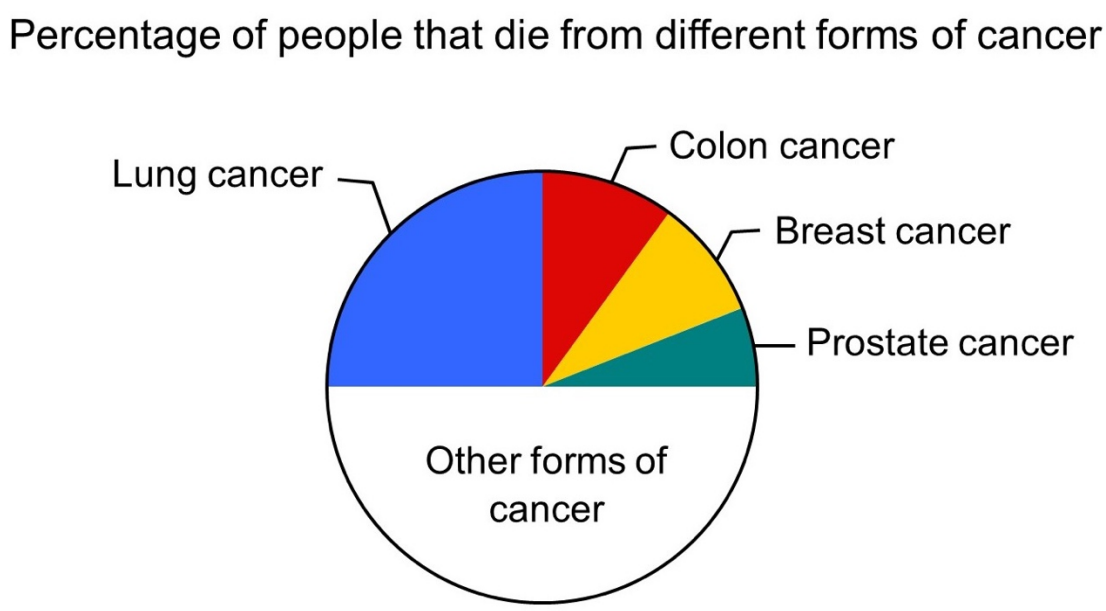
End of Block: Default Question Block

APPENDIX E

SHORT GRAPHICAL LITERACY SCALE

Start of Block: Default Question Block

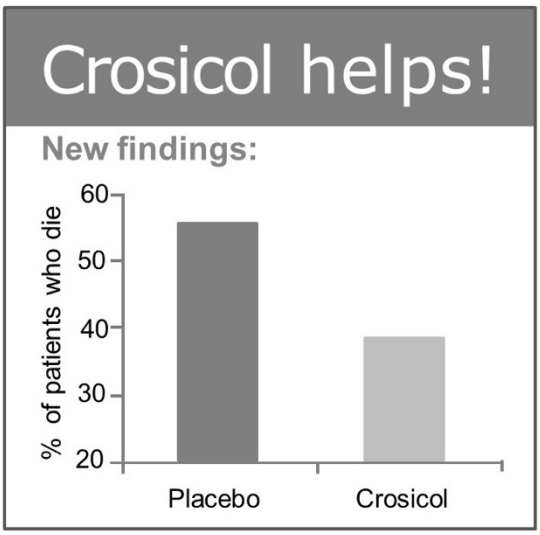
Q3 The graph below shows the percentage of people who die from different types of cancer.

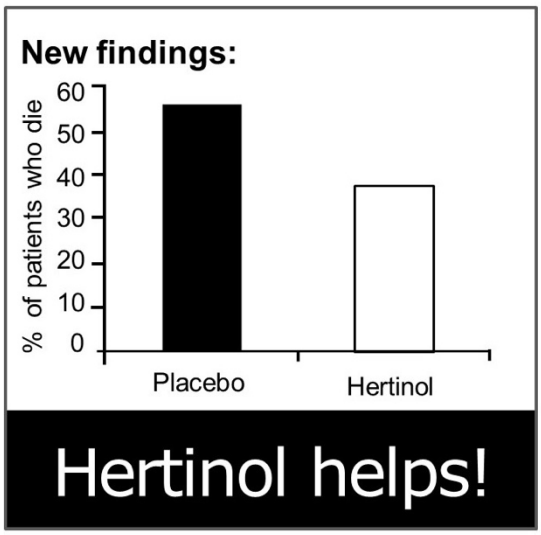


Q4 About what percentage of people who die from cancer die from colon cancer, breast cancer, and prostate cancer taken together? Please enter a number and percent sign (%) below.

_____

Q5 You see two magazine advertisements on separate pages. Each advertisement is for a different drug for treating heart disease. Each advertisement has a graph showing the effectiveness of the drug compared to a placebo (sugar pill).

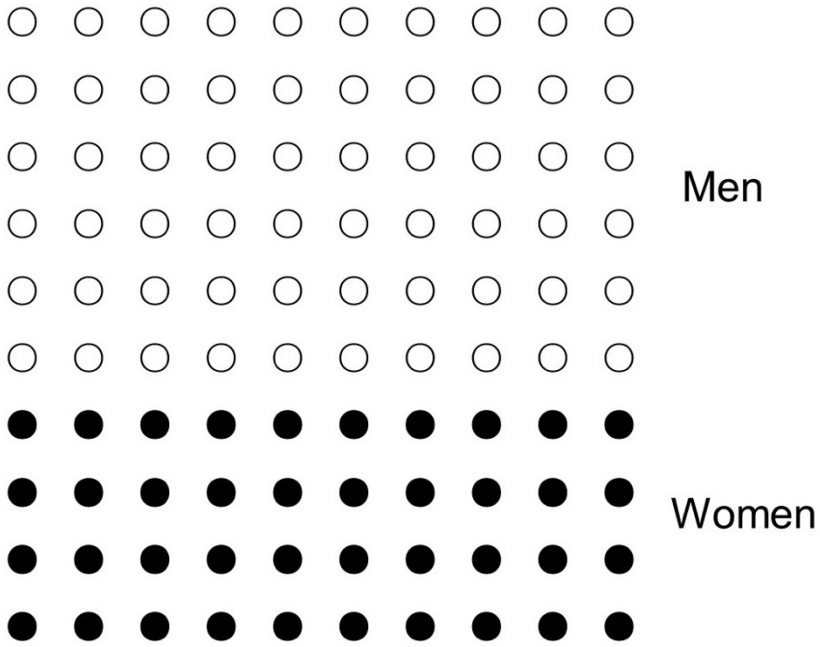




Q6 Compared to the placebo, which treatment leads to a larger decrease in the percentage of patients who die?

- Crosicol (1)
- Hertinol (2)
- They are equal (3)
- Can't say (4)

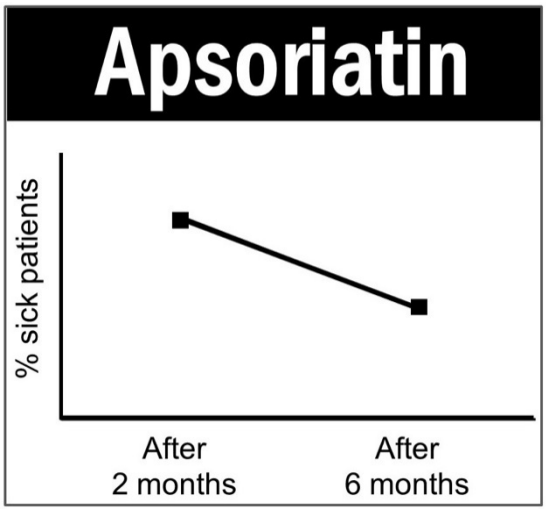
Q7 The graph below shows the number of men and women with disease X. The total numbers of circles is 100.

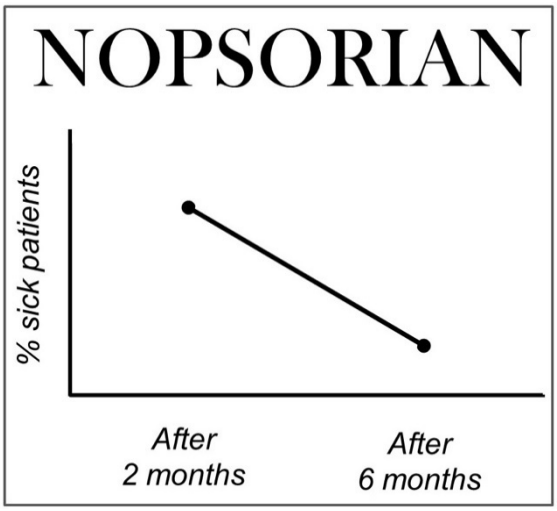


Q8 How many more men than women are there among 100 patients with disease X? Please enter a number below.

_____

Q9 You see two newspaper advertisements on separate pages. Each advertisement is for a different treatment of a skin disease. Each advertisement has a graph showing the effectiveness of the treatment over time.





Q10 Which of the treatments shows a larger decrease in the percentage of sick patients?

- Apsoriatin (1)
- Nopsorian (2)
- They are equal (3)
- Can't say (4)

End of Block: Default Question Block

APPENDIX F

TASK DATA DIFFICULTY CONTROL STRATEGIES

Task	Task Question	Correct Answer (Fitbit)	Correct Answer (Strava)	Correct Answer (Prototype)	Strategy
1	During which month was Jane most active?	June (113 mi)	July (99 mi)	June (113 mi)	Most active month always roughly 10% higher than 2nd highest month for a consistent relative comparison
2	During which month was Jane least active?	December	April	November	Least active month always roughly 20% lower than 2nd lowest month for a consistent relative comparison
3	In June, on which two days of the week did Jane tend to be most active?	Saturday, Sunday	Friday, Sunday	Friday, Saturday	2nd most active day of the week always roughly 20%-30% higher than 3rd lowest day of the week for a consistent relative comparison
4	In June, on which three days of the week did Jane tend to be least active?	Monday, Wednesday, Thursday	Tuesday, Wednesday, Thursday	Monday, Tuesday, Wednesday	3rd least active day of the week always roughly 10% lower than 4th lowest day of the week for a consistent relative comparison
5	What was the date of Jane's most active day in July?	July 29th	July 27th	July 23rd	Most active day in July always 5-10% higher than 2nd most active day in July for a consistent relative comparison
6	What was the date of Jane's least active day in July?	July 7th	July 12th	July 5th	Least active day in July always 15-25% lower than 2nd least active day in July

Task	Task Question	Correct Answer (Fitbit)	Correct Answer (Strava)	Correct Answer (Prototype)	Strategy
					for a consistent relative comparison
7	How far did Jane walk in the entire month of February? You may round to the nearest whole mile.	61 (61.1)	71 (70.78)	59 (59.4)	Keep values in a generally similar range (+/- 25%), do not set February as the extreme high or low in questions 1 or 2 to avoid participants looking up same value twice
8	How far did Jane walk during the week of November 5th? You may round to the nearest whole mile.	11 (11.03)	12 (11.69)	11 (11.01)	Keep values in a generally similar range (+/- 10%)
9	How far did Jane walk on May 12th? You may round to the nearest whole mile.	4 (4.2)	3 (2.5)	4 (4.27)	Keep values away from extreme ends of mile spectrum
10	In September, on how many days did Jane walk less than 1 mile?	8	9	6	Provide 2-3 days within between 0.9 to 1.1 miles
11	At some point in 2018, Jane missed collecting data for 9 days in a row. What is the start date of this 9-day streak?	December 7th	December 18th	November 20th	Place towards end of year since participants will always begin this question in the month of September; if participants have a consistent strategy to check either towards January or December then this will limit variability to time to scan each month in that direction

APPENDIX G

FORMATS OF STANDARD EMAILS SENT TO PARTICIPANTS

Initial Contact Email

Hi [potential participant name],

If contacting through Rare Patient Voice
I'm Peter Frackleton - a master's candidate and researcher at the University of Massachusetts Amherst. I am working with Rare Patient Voice (Project Number UMASS_5378) to recruit participants for my thesis study. They provided me with your contact information following indication from you that you might like to participate and that you meet the requirements for participating in this study.

All other participant leads
I'm Peter Frackleton - a master's candidate and researcher at the University of Massachusetts Amherst. I am currently recruiting participants for my thesis study. I would like to invite you to participate in this study, the details for which can be found below.

About This Study
The main topic of this study is physical activity data collected from fitness tracking devices such as Fitibit and Apple Watch. More specifically, this study will focus on how such data is presented in the form of graphs/data visualizations.

Your participation in this study will be completed in a single session taking roughly 90 minutes to 2 hours. During this session, you will use three different data visualizations showing physical activity data to answer a series of questions. I'll also be asking you about your experiences using these graphs in an interview format and there will be some short questionnaires at the end of our session to collect some additional information. All these steps will be done using Zoom video chat.

For your participation in the study, you will be compensated $40 in the form of an Amazon egift card sent through email.

Recap of Requirements to Participate
To restate the requirements of this study, you must meet the following conditions to be eligible to participate:

1. Be 55+ years old
2. Have a working computer connected to internet (PC or Mac okay, iPad/tablet or iPhone/smartphone will not work for this study)
3. Be able to use Zoom video chat well enough to enter call and view shared screen

If that all sounds good to you and you are still interested in participating, I would be happy to include you as a participant in my study. Please let me know what timing works well for you to set up a Zoom call. I can work around your schedule, so let me know what works best for you.

If you have any questions or concerns about how the study is being conducted or what is required of you as a participant, please feel free to send me any questions via email.

Best regards

Zoom Session Confirmation Email

Hi [participant name],

I set up a link to that Zoom call below for [date, time, time zone]. I have provided a link below. If you want to move the meeting to be any earlier or later, please let me know and I will be happy to accommodate.

Topic: Peter Frackleton's Zoom Meeting
Time: [date, time, time zone]

Join Zoom Meeting
[insert link to Zoom call]

I also prepared a consent form that outlines the basic structure of the study as well as other important information you should be aware of before participating. I have linked it below. You are free to review this at any time and sign either before our session or wait until our session begins to ask any questions and you can sign at that time instead.

https://umassamherst.co1.qualtrics.com/jfe/form/SV_1Bn7fuv6EJOyeTc

Lastly, just to confirm, you have a computer that can be used for Zoom. Correct? Tablets and mobile devices do not work for this study since you will need to interact with things that I will share on my screen. I can explain further at the beginning of the session.

Thank you

APPENDIX H

ZOOM CALL SESSION INTRODUCTORY DEBRIEFING NOTES

The following notes are read to the participant at the beginning of the session prior to the participant interacting with the data visualization and completing the first set of task questions.

- Have you (the participant) already read, understood, and signed the consent form? If not, please review it now and sign when you are ready. Please ask any questions you may have about the contents.
- Please keep in mind that you (the participant) are free to terminate your involvement in the study at any point with no consequence or repercussion.
- During this session, I (the researcher) will be recording audio, screen share video, and audio transcription throughout the call. May I begin recording now?
- Important notes about the data and tasks that we will work with during the session:
    - Will be working with artificial exercise data of fake profile: Jane, meant to simulate typical tracking behavior including regular exercise schedule & lapsing. This means data is not collected by an actual person using a physical activity tracking device.
    - You will work with 3 different data visualizations throughout the duration of this study.
    - All 3 data visualizations contain only one year's worth of data (2018) and only presents data in the form of miles walked.
    - I will ask 11 questions about the data displayed in each data visualization. These questions will remain the same for every data visualization.
    - Each data visualization's data is different, meaning that the answers will not remain the same for the same questions.
    - I (the researcher) cannot not provide guidance regarding interpretation of the information on the graph, only technical questions such as how to navigate visualization interfaces or any problems with Zoom.
    - Some questions may be difficult to answer; this is normal & expected and is nothing to be discouraged about. If the question feels too difficult, you (the participant) may choose to guess or skip for that question.
    - After each of the sets of 11 questions, I will have you (the participant) complete a small questionnaire about how easy or difficult each question was to answer, and then we will have a brief interview about your experience using that data visualization.
- You (the participant) are welcome to get a pencil and paper to have with you as we proceed with our session in case you would like to take notes about the tasks or questions. However, this is not required.
- You are welcome to take a break at any point during the session for any reason.

APPENDIX I

TASK RESPONSE CORRECTNESS DATA

Table I1

Results of chi square tests for incorrect responses between visualizations

Task	Incorrect Responses			Chi Square Test P Value		
	Fitbit (n=30)	Strava (n=30)	Prototype (n=30)	Fitbit - Prototype	Strava - Prototype	Strava - Fitbit
Task 1	5	5	0	6.17E-02	6.17E-02	1.00E+00
Task 2	4	18	1	3.50E-01	**8.98E-06**	**4.96E-04**
Task 3	3	2	0	2.36E-01	4.72E-01	1.00E+00
Task 4	6	3	19	**1.68E-03**	**5.86E-05**	4.70E-01
Task 5	6	0	3	4.70E-01	2.36E-01	**3.14E-02**
Task 6	11	0	13	7.92E-01	**1.70E-04**	**8.49E-04**
Task 7	1	14	20	**1.10E-06**	1.93E-01	**3.47E-04**
Task 8	3	2	1	6.05E-01	1.00E+00	1.00E+00
Task 9	3	1	26	**1.32E-08**	**4.72E-10**	6.05E-01
Task 10	27	9	29	6.05E-01	**3.58E-07**	**7.47E-06**
Task 11	0	7	0	1.00E+00	**1.58E-02**	**1.58E-02**
All Tasks	69	61	112	**2.48E-04**	**9.63E-06**	4.93E-01

APPENDIX J

TASK RESPONSE TIME DATA

Table J1

Results of Tukey HSD for task time between visualizations

Task	Incorrect Responses			ANOVA Test P Value		
	Fitbit (n=30)	**Strava (n=30)**	**Prototype (n=30)**	**Fitbit - Prototype**	**Strava - Prototype**	**Strava - Fitbit**
1	41.5 (39.8)	182.0 (171.0)	14.8 (57.2)	6.01E-01	**9.85E-08**	**5.79E-06**
2	12.1 (11.8)	59.7 (46.7)	14.1 (59.5)	9.83E-01	**3.91E-04**	**2.08E-04**
3	63.3 (40.1)	53.4 (57.2)	27.0 (22.8)	**3.73E-03**	**4.68E-02**	6.37E-01
4	40.0 (39.1)	19.9 (10.6)	18.4 (18.0)	**4.36E-03**	9.71E-01	**8.75E-03**
5	42.4 (30.6)	36.0 (23.1)	16.5 (15.3)	**1.83E-04**	**5.99E-03**	5.52E-01
6	29.2 (28.5)	20.6 (10.8)	21.0 (22.0)	3.13E-01	9.97E-01	2.81E-01
7	23.6 (24.8)	89.3 (47.3)	41.2 (36.5)	1.67E-01	**9.05E-06**	**3.79E-09**
8	55.8 (94.9)	21.5 (11.1)	16.6 (11.6)	**2.08E-02**	9.38E-01	**4.98E-02**
9	41.1 (27.2)	19.3 (13.3)	40.9 (22.2)	9.99E-01	**6.37E-04**	**5.46E-04**
10	41.5 (27.4)	58.3 (25.1)	35.6 (22.9)	6.33E-01	**2.15E-03**	**3.07E-02**
11	58.0 (46.6)	69.0 (72.9)	25.5 (21.5)	**4.32E-02**	**4.30E-03**	6.86E-01
All Tasks	57.2 (77.2)	40.8 (44.3)	24.7 (33.1)	**5.10E-04**	**1.39E-13**	**3.65E-04**

APPENDIX K

TASK DIFFICULTY TALLIES

Table K1

Tally of task difficulty ratings by task/visualization

		Task											
Visualization	**Difficulty**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**Total**
Fitbit	Very Easy	18	22	13	11	15	16	21	22	21	18	20	197
	Somewhat Easy	8	3	8	10	13	12	5	5	4	8	6	82
	Somewhat Difficult	4	5	6	7	2	2	4	3	5	3	3	44
	Very Difficult	0	0	3	2	0	0	0	0	0	1	1	7
Strava	Very Easy	5	9	11	10	20	23	7	18	25	11	18	157
	Somewhat Easy	15	13	17	17	7	5	10	5	2	12	3	106
	Somewhat Difficult	5	5	1	2	3	2	12	6	2	4	3	45
	Very Difficult	5	3	1	1	0	0	1	1	1	3	6	22
Prototype	Very Easy	23	26	13	11	14	12	13	21	6	7	25	171
	Somewhat Easy	7	4	15	16	8	7	9	8	7	6	5	92
	Somewhat Difficult	0	0	2	1	4	4	7	1	7	11	0	37
	Very Difficult	0	0	0	2	4	7	1	0	10	6	0	30

APPENDIX L

ASSIGNED GROUP TASK PERFORMANCE DATA

Table L1

Tally of incorrect task responses and incorrect response rates by task per group and chi square p-value for factor influence

	Task										
Group	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Group 1 (n = 10)	5 (17%)	10 (33%)	4 (13%)	10 (33%)	4 (13%)	7 (23%)	14 (47%)	2 (7%)	11 (37%)	21 (70%)	3 (10%)
Group 2 (n = 10)	1 (3%)	3 (10%)	1 (3%)	9 (30%)	3 (10%)	10 (33%)	10 (33%)	3 (10%)	9 (30%)	21 (70%)	3 (10%)
Group 3 (n = 10)	4 (13%)	10 (33%)	0 (0%)	9 (30%)	2 (7%)	7 (23%)	11 (37%)	1 (3%)	10 (33%)	23 (77%)	1 (3%)
P value	2.32 E-01	5.72 E-02	6.37 E-02	9.49 E-01	6.90 E-01	6.00 E-01	5.45 E-01	5.85 E-01	8.61 E-01	8.01 E-01	5.38 E-01

Table L2

Tally of incorrect task responses and incorrect response rates by visualization per group and chi square p-value for factor influence

	Visualization			
Group	**Fitbit**	**Strava**	**Prototype**	**Total**
Group 1 (n = 10)	31 (28%)	24 (22%)	36 (33%)	91/330 (28%)
Group 2 (n = 10)	16 (15%)	18 (16%)	39 (35%)	73/330 (22%)
Group 3 (n = 10)	22 (20%)	19 (17%)	37 (34%)	78/330 (24%)
P value	**4.36E-02**	5.36E-01	9.10E-01	2.43E-01

Table L3

Mean task time and standard deviation by task per group and ANOVA p-value for factor influence

Group	Task										
	1	2	3	4	5	6	7	8	9	10	11
Group 1 (n = 10)	70.1 (94.1)	37.4 (69.8)	46.6 (34.0)	31.0 (38.0)	33.3 (26.2)	24.2 (21.3)	59.0 (51.7)	31.0 (30.5)	38.2 (31.4)	46.7 (28.2)	54.0 (56.8)
Group 2 (n = 10)	88.8 (130)	26.6 (39.6)	48.3 (39.2)	26.7 (23.2)	27.4 (26.0)	24.3 (20.0)	48.8 (44.3)	18.9 (9.56)	29.4 (17.3)	47.9 (29.2)	50.3 (66.7)
Group 3 (n = 10)	79.6 (158)	21.8 (28.2)	48.8 (58.5)	20.6 (15.1)	34.3 (26.2)	22.4 (24.6)	46.2 (43.0)	44.1 (94.0)	33.6 (20.3)	40.9 (22.7)	48.1 (36.1)
P value	8.57 E-01	4.56 E-01	9.79 E-01	3.30 E-01	5.47 E-01	9.32 E-01	5.32 E-01	2.40 E-01	3.67 E-01	5.66 E-01	9.14 E-01

Table L4

Mean task time and standard deviation by visualization per group and ANOVA p-value for factor influence

Group	Visualization			Across Vizes	Visualization			Across Sums	Total
	F	S	P		F	S	P		
Group 1 (n = 10)	41.6 (41.9)	57.9 (56.3)	29.1 (46.1)	42.9 (49.8)	458 (220)	637 (258)	320 (224)	472 (262)	1415 (655)
Group 2 (n = 10)	37.0 (33.1)	60.1 (80.3)	22.2 (25.5)	39.8 (54.4)	407 (158)	661 (248)	245 (101)	437 (246)	1312 (427)
Group 3 (n = 10)	43.7 (55.2)	53.7 (91.3)	22.7 (22.2)	40.0 (64.0)	481 (198)	591 (419)	250 (109)	440 (302)	1321 (677)
P value	5.17E-01	8.25E-01	2.25E-01	7.36E-01	6.87E-01	8.83E-01	4.83E-01	8.64E-01	9.14E-01

APPENDIX M

TASK PERFORMANCE BY DEMOGRAPHIC CHARACTERISTIC ROUP

Table M1

Number of incorrect responses by task per demographic factor and chi sq p-value for factor influence

Baseline Characteristic	Task										
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Education											
Some College (n = 2)	0 (0%)	2 (33%)	1 (17%)	0 (0%)	1 (17%)	2 (33%)	3 (50%)	1 (17%)	2 (33%)	5 (83%)	1 (17%)
Associate's Degree (n = 3)	2 (22%)	2 (22%)	0 (0%)	1 (11%)	0 (0%)	0 (0%)	4 (44%)	0 (0%)	3 (33%)	6 (67%)	1 (11%)
Bachelor's Degree (n = 10)	1 (3%)	7 (23%)	3 (10%)	14 (47%)	4 (13%)	9 (30%)	11 (37%)	2 (7%)	11 (37%)	21 (70%)	4 (13%)
Master's Degree (n = 12)	6 (17%)	9 (25%)	1 (3%)	9 (25%)	4 (11%)	12 (33%)	11 (31%)	2 (6%)	12 (33%)	26 (72%)	1 (3%)
Professional Degree (n = 3)	1 (11%)	3 (33%)	0 (0%)	4 (44%)	0 0%)	1 (11%)	6 (67%)	1 (11%)	2 (22%)	7 (78%)	0 (0%)
P value	3.04 E-01	9.62 E-01	3.89 E-01	5.74 E-02	6.06 E-01	2.39 E-01	3.47 E-01	7.42 E-01	9.57 E-01	9.49 E-01	3.92 E-01
Career Field											
Business (n = 11)	2 (6%)	7 (21%)	1 (3%)	11 (33%)	3 (9%)	10 (30%)	11 (33%)	2 (6%)	10 (30%)	25 (76%)	3 (9%)
Education (n = 5)	2 (13%)	8 (53%)	1 (7%)	6 (40%)	1 (7%)	5 (33%)	6 (40%)	1 (7%)	6 (40%)	10 (67%)	1 (7%)
Health and medicine (n = 6)	3 (17%)	2 (11%)	0 (0%)	6 (33%)	2 (11%)	2 (11%)	7 (39%)	1 (6%)	4 (22%)	11 (61%)	1 (6%)
Law (n = 1)	1 (33%)	2 (67%)	1 (33%)	2 (67%)	0 (0%)	0 (0%)	2 (67%)	0 (0%)	2 (67%)	3 (100%)	0 (0%)
Public and social services (n = 4)	2 (17%)	3 (25%)	1 (8%)	1 (8%)	2 (17%)	4 (33%)	5 (42%)	1 (8%)	5 (42%)	11 (92%)	1 (8%)
Science, math, and technology (n = 2)	0 (0%)	1 (17%)	0 (0%)	1 (17%)	1 (17%)	2 (33%)	3 (50%)	0 (0%)	2 (33%)	3 (50%)	0 (0%)
Social sciences (n = 1)	0 (0%)	0 (0%)	1 (33%)	1 (33%)	0 (0%)	1 (33%)	1 (33%)	1 (33%)	1 (33%)	2 (67%)	1 (33%)

Baseline Characteristic	Task										
	1	2	3	4	5	6	7	8	9	10	11
P value	6.11 E-01	5.50 E-02	9.32 E-02	4.30 E-01	9.37 E-01	6.28 E-01	9.37 E-01	6.52 E-01	7.69 E-01	3.71 E-01	7.14 E-01
Gender											
Female (n = 26)	8 (10%)	23 (29%)	5 (6%)	25 (32%)	6 (8%)	19 (24%)	30 (38%)	6 (8%)	27 (35%)	59 (76%)	6 (8%)
Male (n = 4)	2 (17%)	0 (0%)	0 (0%)	3 (25%)	3 (25%)	5 (42%)	5 (42%)	0 (0%)	3 (25%)	6 (50%)	1 (8%)
P value	8.69 E-01	6.80 E-02	8.21 E-01	8.76 E-01	1.79 E-01	3.62 E-01	1.00	7.09 E-01	7.42 E-01	1.34 E-01	1.00

Table M2

Number of incorrect responses by visualization per demographic factor and chi sq p-value for factor influence

Baseline Characteristic	Visualization			Total
	Fitbit	Strava	Prototype	
Education				
Some College (n = 2)	5 (23%)	6 (27%)	7 (32%)	18/66 (27%)
Associate's Degree (n = 3)	4 (12%)	6 (18%)	9 (27%)	19/99 (19%)
Bachelor's Degree (n = 10)	24 (22%)	24 (22%)	39 (35%)	93/330 (26%)
Master's Degree (n = 12)	30 (23%)	18 (14%)	45 (34%)	93/396 (23%)
Professional Degree (n = 3)	6 (18%)	7 (21%)	12 (36%)	25/99 (25%)
P value	7.26 E-01	3.84 E-01	9.25 E-01	6.17 E-01
Career Field				
Business (n = 11)	20 (17%)	22 (18%)	43 (36%)	85/363 (23%)
Education (n = 5)	20 (36%)	7 (13%)	20 (36%)	47/165 (28%)
Health and medicine (n = 6)	10 (15%)	8 (12%)	21 (32%)	39/198 (20%)
Law (n = 1)	5 (45%)	4 (36%)	4 (36%)	13/33 (40%)
Public and social services (n = 4)	9 (20%)	14 (32%)	13 (30%)	36/132 (27%)
Science, math, and technology (n = 2)	4 (18%)	3 (14%)	6 (27%)	13/66 (20%)
Social sciences (n = 1)	1 (10%)	3 (27%)	5 (45%)	9/33 (27%)
P value	**1.54 E-02**	7.62 E-02	9.28 E-01	1.43 E-01
Gender				
Female (n = 26)	60 (21%)	57 (20%)	97 (33%)	214/858 (25%)
Male (n = 4)	9 (20%)	4 (9%)	15 (34%)	28/132 (21%)
P value	1.00	1.30 E-01	1.00	4.13 E-01

Table M3

Mean task time and standard deviation (SD in parentheses) by task per baseline characteristic and ANOVA p-value for factor influence

Baseline Characteristic	Task										
	1	2	3	4	5	6	7	8	9	10	11
Education											
Some College (n = 2)	200 (338)	18.2 (21.8)	94.0 (112)	20.0 (12.5)	42.3 (41.1)	15.1 (11.8)	69.6 (79.4)	105 (207)	41.1 (26.7)	47.5 (19.0)	54.7 (32.5)
Associate's Degree (n = 3)	75.1 (135)	18.7 (23.6)	45.7 (19.3)	19.6 (14.0)	25.0 (10.4)	18.7 (13.3)	40.2 (28.4)	32.7 (31.0)	30.4 (17.4)	43.6 (33.0)	48.4 (59.6)
Bachelor's Degree (n = 10)	76.0 (113)	24.0 (40.7)	53.1 (43.8)	35.2 (41.6)	34.5 (29.8)	26.4 (24.4)	56.8 (50.6)	28.0 (24.0)	42.0 (31.7)	47.3 (26.6)	66.2 (70.5)
Master's Degree (n = 12)	70.3 (86.3)	34.8 (59.8)	40.5 (31.5)	21.9 (14.6)	29.3 (24.7)	24.0 (22.8)	44.9 (35.9)	23.3 (22.8)	25.4 (15.0)	42.1 (25.6)	44.4 (43.0)
Professional Degree (n = 3)	52.3 (66.4)	36.0 (60.7)	31.8 (21.9)	23.1 (15.1)	31.3 (17.1)	23.4 (22.2)	57.6 (58.7)	24.2 (22.1)	37.9 (16.6)	50.4 (32.9)	25.0 (20.9)
P value	2.12 E-01	8.03 E-01	5.77 E-02	2.71 E-01	6.98 E-01	7.68 E-01	6.11 E-01	**2.68 E-02**	5.43 E-02	8.99 E-01	2.87 E-01
Career Field											
Business (n = 11)	87.7 (166)	24.1 (39.7)	61.9 (58.8)	29.0 (36.8)	35.3 (29.2)	26.4 (23.2)	51.6 (46.5)	41.8 (89.6)	36.6 (25.5)	47.3 (26.1)	53.1 (45.6)
Education (n = 5)	75.5 (100)	47.3 (85.9)	37.3 (34.2)	20.6 (11.3)	28.5 (25.1)	20.7 (16.5)	42.9 (40.5)	28.7 (34.7)	26.1 (14.7)	36.6 (17.1)	48.8 (57.3)
Health and medicine (n = 6)	81.0 (117)	30.9 (44.0)	34.8 (30.0)	29.5 (25.2)	30.2 (26.8)	25.3 (27.4)	63.8 (54.3)	28.3 (27.1)	33.0 (18.7)	44.9 (29.5)	46.6 (61.0)
Law (n = 1)	73.3 (109)	41.6 (48.5)	76.2 (21.1)	55.8 (36.3)	42.1 (24.4)	46.1 (31.6)	110 (90.0)	36.9 (7.04)	96.4 (28.8)	91.3 (39.6)	114 (85.3)
Public and social services (n = 4)	59.6 (89.2)	22.0 (33.1)	43.4 (34.4)	22.7 (12.0)	31.8 (26.6)	17.6 (16.5)	40.8 (27.8)	20.6 (10.1)	27.5 (11.2)	46.3 (29.1)	39.6 (30.1)
Science, math, and technology (n = 2)	60.8 (64.3)	14.5 (12.4)	21.3 (16.0)	14.0 (3.69)	25.8 (13.5)	18.8 (7.37)	32.4 (27.1)	13.9 (9.60)	19.7 (11.5)	39.9 (17.0)	16.3 (5.92)
Social sciences (n = 1)	125 (194)	12.4 (17.6)	68.3 (19.0)	9.87 (5.12)	16.8 (13.7)	7.59 (1.49)	36.8 (28.4)	19.6 (6.03)	35.6 (33.8)	26.5 (6.14)	111 (119)
P value	9.89 E-01	7.23 E-01	1.32 E-01	2.96 E-01	8.58 E-01	3.30 E-01	1.98 E-01	9.00 E-01	**5.71 E-05**	**4.68 E-02**	8.23 E-02
Gender											
Female (n = 26)	79.1 (129)	31.3 (51.8)	48.1 (46.3)	25.9 (27.2)	31.7 (25.2)	24.0 (22.3)	51.8 (47.5)	33.9 (61.5)	35.7 (24.7)	45.3 (26.6)	52.5 (51.3)
Male (n = 4)	81.9 (137)	10.7 (14.6)	46.7 (33.3)	27.6 (28.0)	31.5 (31.9)	21.4 (19.5)	48.4 (39.6)	14.6 (8.30)	20.8 (9.78)	44.5 (28.5)	39.8 (71.7)
P value	9.46 E-01	1.76 E-01	9.21 E-01	8.37 E-01	9.77 E-01	7.05 E-01	8.15 E-01	2.83 E-01	**4.25 E-02**	9.30 E-01	4.53 E-01

Table M4

Mean total task time and standard deviation for each visualization (F = Fitbit, S = Strava, P = Prototype) per demographic factor and ANOVA p-value for factor influence

Baseline Characteristic	Visualization			Across Vizes	Visualization			Across Sums	Total
	F	S	P		F	S	P		
Education									
Some College (n = 2)	57.6 (110)	108 (187)	27.1 (26.1)	64.3 (129)	634 (405)	1189 (744)	298 (125)	707 (555)	2121 (1274)
Associate's Degree (n = 3)	32.2 (35.4)	56.1 (72.1)	20.3 (16.4)	36.2 (49.2)	354 (144)	618 (298)	223 (87.1)	398 (244)	1195 (520)
Bachelor's Degree (n = 10)	46.2 (41.2)	62.5 (74.9)	24.8 (23.8)	44.5 (53.4)	508 (184)	688 (264)	273 (129)	490 (259)	1469 (496)
Master's Degree (n = 12)	37.2 (32.4)	47.0 (47.9)	25.1 (42.0)	36.4 (42.1)	410 (179)	517 (208)	276 (198)	401 (214)	1202 (535)
Professional Degree (n = 3)	34.3 (20.5)	47.6 (47.4)	25.3 (36.7)	35.7 (37.4)	377 (83.4)	523 (195)	278 (177)	393 (174)	1179 (447)
P value	1.06 E-01	**1.10 E-02**	9.48 E-01	2.10 E-03	3.48 E-01	**4.96 E-02**	9.87 E-01	8.40 E-02	2.70 E-01
Career Field									
Business (n = 11)	43.5 (57.7)	65.1 (96.3)	26.4 (24.8)	45.0 (68.1)	478 (222)	716 (386)	290 (98.1)	495 (310)	1484 (644)
Education (n = 5)	39.1 (35.7)	46.4 (50.5)	27.1 (59.1)	37.5 (49.7)	430 (204)	511 (285)	298 (300)	413 (263)	1239 (767)
Health and medicine (n = 6)	45.7 (33.9)	53.8 (74.1)	22.8 (28.7)	40.8 (51.3)	503 (149)	592 (277)	251 (121)	448 (235)	1345 (441)
Law (n = 1)	68.9 (59.2)	92.5 (63.7)	52.5 (33.2)	71.3 (54.7)	758 (N/A)	1018 (N/A)	577 (N/A)	784 (221)	2352 (N/A)
Public and social services (n = 4)	33.3 (26.9)	51.0 (50.6)	17.2 (14.9)	33.8 (36.7)	366 (79.2)	561 (167)	189 (25.9)	372 (186)	1116 (243)
Science, math, and technology (n = 2)	22.7 (18.8)	36.9 (36.5)	16.1 (10.6)	25.2 (25.7)	250 (9.07)	406 (55.6)	177 (59.0)	277 (111)	832 (124)
Social sciences (n = 1)	27.6 (24.3)	75.5 (114)	24.8 (27.0)	42.7 (70.7)	304 (N/A)	831 (N/A)	273 (N/A)	469 (313)	1408 (N/A)
P value	6.84 E-02	3.04 E-01	5.30 E-02	3.49 E-03	3.32 E-01	5.87 E-01	4.13 E-01	1.49 E-01	4.41 E-01
Gender									
Female (n = 26)	42.1 (45.9)	57.8 (76.4)	25.3 (34.7)	41.8 (56.7)	463 (178)	636 (309)	279 (164)	459 (267)	1378 (580)
Male (n = 4)	31.9 (31.2)	53.4 (82.8)	20.5 (19.7)	35.3 (53.7)	351 (259)	587 (350)	226 (37.1)	388 (277)	1164 (612)
P value	1.54 E-01	7.24 E-01	3.70 E-01	2.18 E-01	2.77 E-01	7.75 E-01	5.31 E-01	3.94 E-01	5.00 E-01

APPENDIX N

DEMOGRAPHIC COMPOSITION BY ASSIGNED GROUP

Table N1

Distributions of demographics across groups with p value from chi square test to ensure there is no statistically significant difference between groups

Baseline Characteristic	Group 1 (n=10)	Group 2 (n=10)	Group 3 (n=10)	P value
Education				
Some college, no degree	1	0	1	5.85E-01
Associate's degree	1	1	1	1.00
Bachelor's degree	5	4	1	1.43E-01
Master's degree	3	4	5	6.59E-01
Professional degree	0	1	2	3.92E-01
Career field				
Business	4	3	4	8.61E-01
Education	2	1	2	7.87E-01
Health and medicine	1	2	3	5.35E-01
Law	1	0	0	3.56E-01
Public and social services	1	2	1	7.49E-01
Science, math, and technology	1	1	0	5.85E-01
Social sciences	0	1	0	3.56E-01
Gender				
Male	2	2	0	3.15E-01
Female	8	8	10	3.15E-01
Ethnicity				
White	10	10	10	1.00

Table N2

Distributions of additional factors across groups with p value from chi square test or ANOVA to ensure there is no statistically significant difference between groups

Baseline Characteristic	Group 1 (n=10)	Group 2 (n=10)	Group 3 (n=10)	P value
Fitness Tracking Experience				
Has engaged in tracking	8	5	9	1.09E-01
Mean tracking time in years (SD)	3.3 (3.6)	3.6 (2.4)	7.2 (6.3)	2.69E-01
Device on wrist	5	1	5	1.01E-01
Smartphone only	3	2	2	8.30E-01
Other	0	2	2	3.15E-01
Activity				
AVG hrs activity per day	5.4 (3.7)	5.1 (2.8)	6.3 (1.5)	5.79E-01
GLS				
Correct Answers	1.5 (1.2)	2.6 (1.0)	1.9 (0.54)	**1.57E-04**

APPENDIX O

CONTRAST RATIOS BETWEEN FEATURES ON PROTOTYPE VISUALIZATION

Table O

Color values and details on data points being compared for each task as well as resulting contrast ratio between these colors

Task	Low Color	High Color	Contrast Ratio
3	#7ED18B (Sun, 3.8)	#3BBB51 (Fri, 5.8)	1.35:1
4	#8ED69A (Thu, 3.3)	#9ADBA5 (Mon, 2.9)	1.06:1
5	#4AC05E (Jul 20, 5.3 mi)	#3BBB50 (Jul 23, 5.8 mi)	1.06:1
6	#EBF6ED (Jul 5, 0.5)	#E8F5EA (17 + 28, 0.6),	1.01:1,
		#E3F4E6 (Jul 31, 0.8)	1.03:1
9, 10	#FCFCFC (0 mile)	#DAF1DE (1.01 mile)	1.16: 1
	#DAF1DE (1.01 mile)	#B7E5BF (2 mile)	1.17: 1

BIBLIOGRAPHY

[1] Centers for Disease Control and Prevention. Underlying Cause of Death, 1999–2018. CDC WONDER Online Database. Atlanta, GA: Centers for Disease Control and Prevention; 2018. Accessed March 12, 2020.

[2] *Heart Health and Aging*. (2018, June 1). National Institute on Again. https://www.nia.nih.gov/health/heart-health-and-aging

[3] Taylor RS, Sagar VA, Davies EJ, Briscoe S, Coats AJ, Dalal H, Lough F, Rees K, Singh S. Exercise-based rehabilitation for heart failure. Cochrane Database Syst Rev. 2014;(4):CD003331. doi: 10.1002/14651858.CD003331.pub4.

[4] Ito S. (2019). High-intensity interval training for health benefits and care of cardiac diseases - The key to an efficient exercise protocol. *World journal of cardiology*, *11*(7), 171–188. https://doi.org/10.4330/wjc.v11.i7.171

[5] Blair SN, Kampert JB, Kohl HW, et al. Influences of Cardiorespiratory Fitness and Other Precursors on Cardiovascular Disease and All-Cause Mortality in Men and Women. *JAMA.* 1996;276(3):205–210. doi:10.1001/jama.1996.03540030039029

[6] Centers for Disease Control and Prevention.National Center for Chronic Disease Prevention and Health Promotion, Division of Nutrition, Physical Activity, and Obesity.Data, Trend and Maps [online]. [accessed Jun 17, 2021]. URL: https://www.cdc.gov/nccdphp/dnpao/data-trends-maps/index.html

[7] Brickwood K, Watson G, O'Brien J, Williams A, Consumer-Based Wearable Activity Trackers Increase Physical Activity Participation: Systematic Review and Meta-

Analysis JMIR Mhealth Uhealth 2019;7(4):e11819 URL: https://mhealth.jmir.org/2019/4/e11819 DOI: 10.2196/11819

[8] Burke L.E., Conroy M.B., Sereika S.M., Elci O.U., Styn M.A., Acharya S.D., Sevick M.A., Ewing L.J., Glanz K. The Effect of Electronic Self-Monitoring on Weight Loss and Dietary Intake: A Randomized Behavioral Weight Loss Trial. Obesity. 2011;19:338–344. doi: 10.1038/oby.2010.208

[9] King A.C., Ahn D.K., Oliveira B.M., Atienza A.A., Castro C.M., Gardner C.D. Promoting Physical Activity Through Hand-Held Computer Technology. Am. J. Prev. Med. 2008;34:138–142. doi: 10.1016/j.amepre.2007.09.025.

[10] King A.C., Hekler E.B., Grieco L.A., Winter S.J., Sheats J.L., Buman M.P., Banerjee B., Robinsom T.N., Cirimele J. Effects of Three Motivationally Targeted Mobile Device Applications on Initial Physical Activity and Sedentary Behavior Change in Midlife and Older Adults: A Randomized Trial. PLoS ONE. 2016;11:e0156370. doi: 10.1371/journal.pone.0156370.

[11] Van Het R.E., Silveira P., van de L.R., Daniel F., Casati F., de Bruin E.D. Tablet-Based Strength-Balance Training to Motivate and Improve Adherence to Exercise in Independently Living Older People: Part 2 of A Phase II Preclinical Exploratory Trial. J. Med. Internet Res. 2016;18:e5. doi: 10.2196/jmir.3055.

[12] Epstein, D. A., Kang, J. H., Pina, L. R., Fogarty, J., & Munson, S. A. (2016). Reconsidering the device in the drawer: Lapses as a design opportunity in personal informatics. *UbiComp 2016 - Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 829–840. https://doi.org/10.1145/2971648.2971656

[13] Daniel A. Epstein, Monica Caraway, Chuck Johnston, An Ping, James Fogarty, and Sean A. Munson. (2016). Beyond Abandonment to Next Steps: Understanding and Designing for Life after Personal Informatics Tool Use. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2016)*, 1109-1103.

[14] McMahon S, Lewis B, Oakes M, Guan W, Wyman J, Rothman A (2016). Older Adults' Experiences Using a Commercially Available Monitor to Self-Track Their Physical Activity. JMIR Mhealth Uhealth 2016;4(2):e35. doi: 10.2196/mhealth.5120

[15] Sullivan, A. N., & Lachman, M. E. (2017). Behavior Change with Fitness Technology in Sedentary Adults: A Review of the Evidence for Increasing Physical Activity. Frontiers in public health, 4, 289. https://doi.org/10.3389/fpubh.2016.00289

[16] HHS Office of the Secretary,Office for Civil Rights, and Ocr. "Health Information Technology." *HHS.gov*, US Department of Health and Human Services, 19 Apr. 2019, www.hhs.gov/hipaa/for-professionals/special-topics/health-information-technology/index.html.

[17] "American Recovery and Reinvestment Act - ARRA." *HITECH Answers: HIPAA, MIPS, EHR, Cybersecurity News*, 2019, www.hitechanswers.net/about/about-arra/.

[18] HHS Office of the Secretary,Office for Civil Rights, and Ocr. "HITECH Act Enforcement Interim Final Rule." *HHS.gov*, US Department of Health and Human Services, 16 June 2017, www.hhs.gov/hipaa/for-professionals/special-topics/hitech-act-enforcement-interim-final-rule/index.html.

[19] “Introduction | Meaningful Use | CDC.” *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 9 Sept. 2019, www.cdc.gov/ehrmeaningfuluse/introduction.html.

[20] “Health IT Playbook.” *HealthIT.gov*, 2019, www.healthit.gov/playbook/certified-health-it/#section-2-1.

[21] Patel, Vaishali, et al. “Trends in Consumer Access and Use of Electronic Health Information.” *HealthIT.gov*, 15 Oct. 2015, ii. https://www.healthit.gov/sites/default/files/playbook/pdf/data-brief-trends-in-consumer-access-and-use-of-electronic-health-info.pdf.

[22] Gheorghiu, B., & Hagens, S. (2017). Use and maturity of electronic patient portals. *Studies in Health Technology and Informatics*, *234*, 136–141. https://doi.org/10.3233/978-1-61499-742-9-136

[23] Saripalle, R. (2019). Integrating physical activity data with electronic health record. *HEALTHINF 2019 - 12th International Conference on Health Informatics, Proceedings; Part of 12th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2019*, 21–29.

[24] Van Doornik, William. “Meaningful Use of Patient-Generated Data in EHRs.” *Journal of AHIMA*, Oct. 2013, library.ahima.org/doc?oid=106996#.XZZW-C2ZPR2.

[25] Cohen, D. J., Keller, S. R., Hayes, G. R., Dorr, D. A., Ash, J. S., & Sittig, D. F. (2016). Integrating Patient-Generated Health Data Into Clinical Care Settings or Clinical Decision-Making: Lessons Learned From Project HealthDesign. *JMIR Human Factors*, *3*(2), e26. https://doi.org/10.2196/humanfactors.5919

[26] Dinh-Le, C., Chuang, R., Chokshi, S., & Mann, D. (2019). Wearable Health Technology and Electronic Health Record Integration: Scoping Review and Future Directions. *JMIR MHealth and UHealth*, *7*(9), e12861. https://doi.org/10.2196/12861

[27] Page, T. (2018). A forecast of the adoption of wearable technology. *Wearable Technologies: Concepts, Methodologies, Tools, and Applications*, 1370–1388. https://doi.org/10.4018/978-1-5225-5484-4.ch063

[28] Fox, S., & Duggan, M. (2013). Tracking for Health. *Pew Internet*, (January), 1–40. https://doi.org/10.1001/jamainternmed.2013.1221.2.

[29] Choe, E. K., Lee, N. B., Lee, B., Pratt, W., & Kientz, J. A. (2014). Understanding quantified-selfers' practices in collecting and exploring personal data. *Conference on Human Factors in Computing Systems - Proceedings*, 1143–1152. https://doi.org/10.1145/2556288.2557372

[30] Li, I., Dey, A., & Forlizzi, J. (2010). A stage-based model of personal informatics systems. *Conference on Human Factors in Computing Systems - Proceedings*, *1*, 557–566. https://doi.org/10.1145/1753326.1753409

[31] Li, I., Dey, A. K., & Forlizzi, J. (2011). Understanding my data, myself: Supporting self-reflection with ubicomp technologies. *UbiComp'11 - Proceedings of the 2011 ACM Conference on Ubiquitous Computing*, 405–414. https://doi.org/10.1145/2030112.2030166

[32] Fawcett, T. (2015). Mining the quantified self: Personal knowledge discovery as a challenge for data science. *Big Data*, *3*(4), 249–266. https://doi.org/10.1089/big.2015.0049

[33] Choe, E. K., Lee, B., & Schraefel, M. C. (2015). Characterizing Visualization Insights from Quantified Selfers' Personal Data Presentations. *IEEE Computer Graphics and Applications*, *35*(4), 28–37. https://doi.org/10.1109/MCG.2015.51

[34] "Chapter 1 – Our Backbone: Why we Visualize." *Effective Data Visualization: the Right Chart for the Right Data*, by Stephanie D. H. Evergreen, SAGE, 2017, pp. 1-6.

[35] "Visualizing Time Series Data." *Data Points: Visualization That Means Something*, by Nathan Yau, John Wiley & Sons, Inc., 2013, pp. 154–165.

[36] "Chapter 5 - Variation Through Time." *Signal: Understanding What Matters in a World of Noise*, by Stephen Few, Analytics Press, 2015, pp. 105–115.

[37] Hossain, A., & Zaman, T. (2012). AC 2012-3605 : HMI DESIGN : AN ANALYSIS OF A GOOD DISPLAY FOR SEAMLESS INTEGRATION BETWEEN USER UNDERSTANDING AND HMI Design : An Analysis of a Good Display for Seamless Integration Between User Understanding and Automatic Controls Abstract : In process a. *American Society for Engineering Education*, 14.

[38] Cawthon, N., & Moere, A. Vande. (2007). The effect of aesthetic on the usability of data visualization. *Proceedings of the International Conference on Information Visualisation*, 637–645. https://doi.org/10.1109/IV.2007.147

[39] Chittaro, L. "Visualizing Information on Mobile Devices." *Computer*, vol. 39, no. 3, 2006, pp. 40–45., doi:10.1109/mc.2006.109.

[40] Games, P. S., & Joshi, A. (2013). Visualization of off-screen data on tablets using context-providing bar graphs and scatter plots. *Visualization and Data Analysis 2014*, *9017*(February 2014), 90170D. https://doi.org/10.1117/12.2038456

[41] Fan, C., Forlizzi, J., & Dey, A. K. (2012). A spark of activity: Exploring informative art as visualization for physical activity. *UbiComp'12 - Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, 81–84. https://doi.org/10.1145/2370216.2370229

[42] Choe, E. K., Lee, B., Zhu, H., & Riche, N. H. (2017). Understanding self-reflection: How people reflect on personal data through visual data exploration. *ACM International Conference Proceeding Series*, (October), 173–182. https://doi.org/10.1145/3154862.3154881

[43] Liang, Z., Ploderer, B., Liu, W., Nagata, Y., Bailey, J., Kulik, L., & Li, Y. (2016). SleepExplorer: a visualization tool to make sense of correlations between personal sleep data and contextual factors. *Personal and Ubiquitous Computing*, *20*(6), 985–1000. https://doi.org/10.1007/s00779-016-0960-6

[44] Pavel, D., Trossen, D., Holweg, M., & Callaghan, V. (2013). Lifestyle stories: Correlating user information through a story-inspired paradigm. *Proceedings of the 2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops, PervasiveHealth 2013*, (May), 412–415. https://doi.org/10.4108/icst.pervasivehealth.2013.252131

[45] Fry, E. (1981). Graphical Literacy. *Journal of Reading, 24*(5), 383-389. Retrieved from http://www.jstor.org/stable/40032373

[46] Okan, Y., Galesic, M., & Garcia-Retamero, R. (2016). *How People with Low and High Graph Literacy Process Health Graphs : Evidence from Eye-tracking . White Rose Research Online URL for this paper : Article : Okan , Y , Galesic , M and Garcia-Retamero , R (2016) How People with Low and High Graph Literacy*. *29*,

271–294. Retrieved from file:///Users/Moritz/Desktop/Master Seminar/Literature/Okanetal2015_JBDM_author.pdf

[47] Rodríguez, V., Andrade, A. D., García-Retamero, R., Anam, R., Rodríguez, R., Lisigurski, M., … Ruiz, J. G. (2013). Health literacy, numeracy, and graphical literacy among veterans in primary care and their effect on shared decision making and trust in physicians. *Journal of Health Communication*, *18*(SUPPL. 1), 273–289. https://doi.org/10.1080/10810730.2013.829137

[48] Nelson, W., Reyna, V. F., Fagerlin, A., Lipkus, I., & Peters, E. (2008). Clinical implications of numeracy: Theory and practice. *Annals of Behavioral Medicine*, *35*(3), 261–274. https://doi.org/10.1007/s12160-008-9037-8

[49] Galesic, M., & Garcia-Retamero, R. (2011). Graph literacy: A cross-cultural comparison. *Medical Decision Making*, *31*(3), 444–457. https://doi.org/10.1177/0272989X10373805

[50] Okan, Y., Janssen, E., Galesic, M., & Waters, E. A. (2019). Using the Short Graph Literacy Scale to Predict Precursors of Health Behavior Change. *Medical Decision Making*, *39*(3), 183–195. https://doi.org/10.1177/0272989X19829728

[51] Tsuji, B. H., & Lindgaard, G. (2014). Comparing novices & experts in their exploration of data in line graphs. *11th International Conference on Cognition and Exploratory Learning in Digital Age, CELDA 2014*, (Celda), 39–46.

[52] Rosenbaum, S., Morell, R., Abdel-Baki, A. *et al.* Assessing physical activity in people with mental illness: 23-country reliability and validity of the simple physical

activity questionnaire (SIMPAQ). *BMC Psychiatry* **20,** 108 (2020). https://doi.org/10.1186/s12888-020-2473-0

[53] Colblindor. (2021). Coblis – color blindness simulator. Retrieved from https://www.color-blindness.com/coblis-color-blindness-simulator/

[54] DEREFELDT, G., LENNERSTRAND, G. and LUNDH, B. (1979), AGE VARIATIONS IN NORMAL HUMAN CONTRAST SENSITIVITY. Acta Ophthalmologica, 57: 679-690. https://doi.org/10.1111/j.1755-3768.1979.tb00517.x

[55] Nielsen, Jakob. “10 Heuristics for User Interface Design: Article by Jakob Nielsen.” *Nielsen Norman Group*, 24 Apr. 1994, www.nngroup.com/articles/ten-usability-heuristics/.

[56] Forsell, C., & Johansson, J. (2010). An heuristic set for evaluation in Information Visualization. *Proceedings of the Workshop on Advanced Visual Interfaces AVI*, (April), 199–206. https://doi.org/10.1145/1842993.1843029

[57] H. Väätäjä, J. Varsaluoma, T. Heimonen, K. Tiitinen, J. Hakulinen, M. Turunen, et al., "Information visualization heuristics in practical expert evaluation", *Proc. Beyond Time Errors Novel Eval. Methods Vis.*, pp. 36-43, 2016, Aug. 12, 2021, [online] Available: https://dl.acm.org/doi/10.1145/2993901.2993918.

[58] Tory, M., & Möller, T. (2005). Evaluating Visualizations : Do Expert Reviews Work ? Graphics, Usability and Visualization Lab When formal user studies fail. *IEEE Computer Graphics and Applications*, *25*(5), 8–11. https://doi.org/10.1016/j.jallcom.2007.01.062

[59] C. Forsell, "Evaluation in Information Visualization: Heuristic Evaluation," 2012 16th International Conference on Information Visualisation, 2012, pp. 136-142, doi: 10.1109/IV.2012.33.

[60] Danny T Y Wu, Annie T Chen, John D Manning, Gal Levy-Fix, Uba Backonja, David Borland, Jesus J Caban, Dawn W Dowding, Harry Hochheiser, Vadim Kagan, Swaminathan Kandaswamy, Manish Kumar, Alexis Nunez, Eric Pan, David Gotz, Evaluating visual analytics for health informatics applications: a systematic review from the American Medical Informatics Association Visual Analytics Working Group Task Force on Evaluation, *Journal of the American Medical Informatics Association*, Volume 26, Issue 4, April 2019, Pages 314–323, https://doi.org/10.1093/jamia/ocy190

[61] McCaffery, K. J., Dixon, A., Hayen, A., Jansen, J., Smith, S., & Simpson, J. M. (2012). The influence of graphic display format on the interpretations of quantitative risk information among adults with lower education and literacy: A 1y. *Medical Decision Making*, *32*(4), 532–544. https://doi.org/10.1177/0272989X11424926

[62] Saraiya, P., North, C., & Duca, K. (2005). An insight-based methodology for evaluating bioinformatics visualizations. *IEEE Transactions on Visualization and Computer Graphics*, *11*(4), 443–456. https://doi.org/10.1109/TVCG.2005.53

[63] Merino, L., Ghafari, M., Anslow, C., & Nierstrasz, O. (2018). A systematic literature review of software visualization evaluation. *Journal of Systems and Software*, Vol. 144, pp. 165–180. https://doi.org/10.1016/j.jss.2018.06.027

[64] Elmqvist, N., & Yi, J. S. (2015). Patterns for visualization evaluation. *Information Visualization*, *14*(3), 250–269. https://doi.org/10.1177/1473871613513228

[65] North, C. “Toward Measuring Visualization Insight.” *IEEE Computer Graphics and Applications*, vol. 26, no. 3, 2006, pp. 6–9., doi:10.1109/mcg.2006.70.

[66] Lam, H., Bertini, E., Isenberg, P., Plaisant, C., Lam, H., Bertini, E., … Seven, S. C. (2012). Seven Guiding Scenarios for Information Visualization Evaluation. *Technical Report No. 2011-992-04 Department of Computer Science University of Calgary*.

[67] Top Overall iOS and Google Play Apps Worldwide. (2019). Apptopia. https://apptopia.com/store-insights/top-charts/itunes-connect/health-fitness/united-states

[68] Rare Patient Voice. (2021). Providing patients and caregivers a voice - Rare Patient Voice. Retrieved from https://rarepatientvoice.com

[69] Brehmer, M., & Munzner, T. (2013). 20190306 -A multi-level typology of abstract visualization tasks-annotated. *IEEE Transactions on Visualization and Computer Graphics*, *19*(12), 2376–2385. https://doi.org/10.1109/TVCG.2013.124

[70] Berinato, S. (2016). *Good Charts: The HBR Guide to Making Smarter, More Persuasive Data Visualizations*. Harvard Business Review Press.

[71] *Color and contrast | Visual design | Accessibility for Teams.* (2021). Accessibility for Teams. https://accessibility.digital.gov/visual-design/color-and-contrast/