AALBORG UNIVERSITY

STUDENT REPORT

Data Visualization in Augmented Reality

Group:

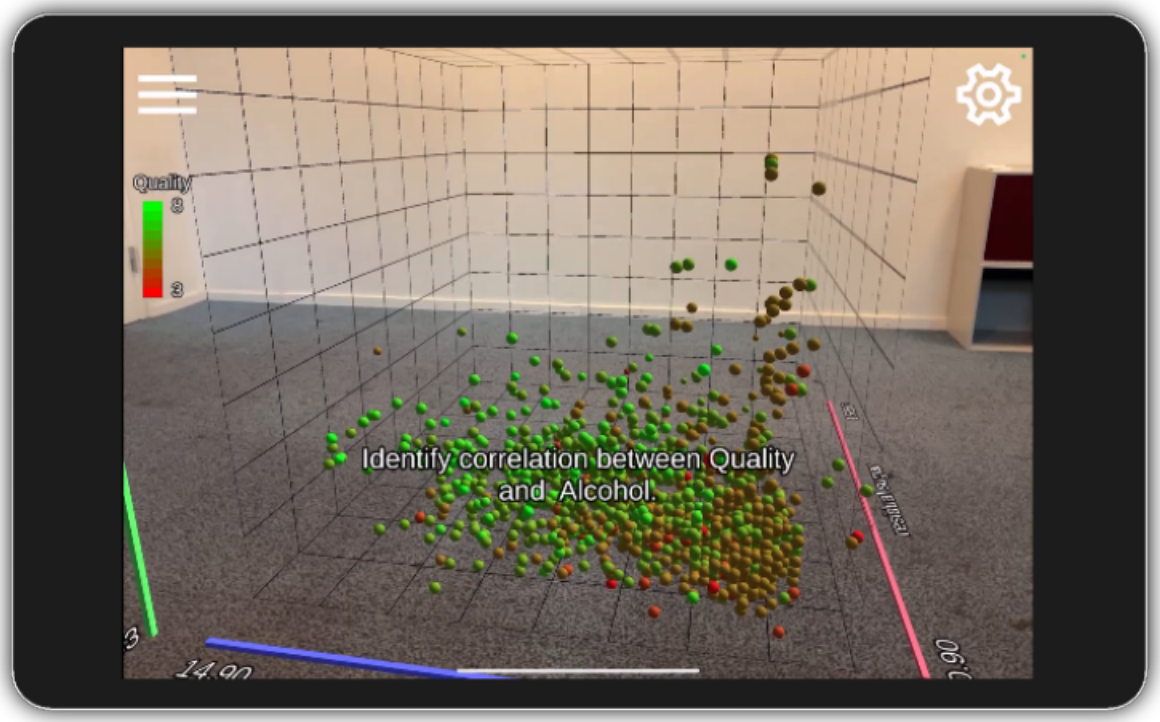
Rasmus K. Schrøder; Stefan L. Olsson; Valdemar B. Petersen

rschra19@student.aau.dk; solsso19@student.aau.dk; vpeter19@student.aau.dk

Supervisor:

Luis Emilio Bruni

leb@create.aau.dk



Number of Pages: 129

Number of Appendices: 26

Completion Date: 24/05-2024

Abstract

This report discusses the findings of a study on data visualization in Augmented Reality (AR). Using frameworks as: The User Experience Questionnaire (UEQ) and NASA-TLX the study explores their relationship with Task Performance metrics such as: Time on Task, Task Success Rate, and Task Duration. Furthermore, the interplay between the frameworks and task performance is examined by physiological measures like: Blood Volume Pulse (BVP) and Electrodermal Activity (EDA). Collectively, these metrics have been used to estimate a value of cognitive load. Developing a dynamic and full AR integrated data visualization application (n = 29) and comparing it to a static and limited AR version (n = 29) provided insights into how user experience and extraneous cognitive load are influenced by the AR medium. This study provides statistically significant findings that user experience in data visualization is positively enhanced, and that extraneous cognitive load is increased, when comparing a full AR integrated data visualization application to a limited one.

Keywords: Data Visualization · Cognitive Load · Augmented Reality · UX Design

Acknowledgements

We, the researchers, would like to express our gratitude to all those who participated in this project. We would like to acknowledge Luis Emilio Bruni who supervised this project. Thanks should also go to the Augmented Cognition Lab at AAU Copenhagen who supported us with physiological equipment and technical guidance.

Table of contents

1 Introduction

In the recent decade, Augmented Reality (AR) has rapidly improved in terms of its technological advancement and popularity, providing new opportunities for creating interactive and immersive experiences [1]. Augmented Reality's potential to enhance data visualization is particularly of interest, as it could offer new and immersive ways for people to visualize, interact, explore, and present data.

We are motivated to explore how augmented reality influences user experience compared to traditional 2D and 3D solutions. We hypothesize that AR data visualization not only creates a more engaging experience but also provides a more straightforward approach to understand simple and complex data visualizations compared to traditional methods. This could be achieved by affording a potentially more intuitive way to control the visualization, i.e. controlling and moving physically. Generally, users must require a certain level of expertise to interpret data visualizations of higher complexities. We hope, with AR and its immersive and intuitive nature, that interpretation of complexities within data visualization could become more accessible to a broader population.

This study will focus on user experience in AR compared to conventional data visualizations. We aim to measure individual participant's user experience through reliable measures to find the overall effectiveness of an AR medium to visualize data through realistic use-cases. This study will undergo a thorough analysis of state-of-the-art technologies and methodologies, development of a prototype and experimental design, and a detailed discussion and conclusion of the findings. The eventual final experiment test should, to the best of our abilities, be based on an optimal prototype, with an encompassing framework, which will attempt to answer a final problem statement, which will be established at a later time. The first step of this process is to establish an Initial Problem Statement.

1.1 Initial Problem Statement

By establishing an Initial Problem Statement (IPS), we have a foundational pillar upon which the project is centered around. Based on the motivation, we establish the IPS as follows:

IPS: *In what ways can augmented reality be used as a medium for enhancing data visualization to optimize user experience?*

2 Analysis

In the Analysis Chapter, we will delve into the current state of the art and relevant literature for augmented reality (AR) and data visualization. This chapter will examine AR technologies, 2D and 3D visualization methods, and important user experience factors, all aimed at enhancing data visualization through AR with a focus on cognitive load.

2.1 Augmented Reality

Augmented reality (AR) is referred to by Carmigniani Et al. as a live view of a physical real-world environment whose elements are merged with augmented computer-generated images, creating a mixed reality [2]. As this augmentation of the real world is typically interactive with a moving device camera, the augmentation is usually done in real time. As the potential of AR has grown phenomenally over the past decade, it has been in our interest to explore AR as a medium or tool in different domains.

Figure 1: The AR feature in Pokémon GO, where Pokémon characters are augmented into the real world for users to interact and catch through the device [3].

Augmented Reality (AR) has been under remarkable development since it was first introduced in the 1960s with the "Sword of Damocles", the first head-mounted display system (HMD) [4]. By the 1990s, huge advancements in modern computer graphics helped introduce augmented reality into aerospace and television, notably in fighter aircraft as HUDs (Heads-Up Displays). In the 2000s there was another significant shift with the introduction of smartphones, which provided an ideal

platform for augmented reality due to their high-quality cameras, GPS, and motion sensors. Pokémon GO was the final catalyst for the popularity of augmented reality due to its global success in 2016, it popularized AR among the general public but also showed what kind of immersive and engaging experiences it could provide.

With evolving technologies such as Apple Vision Pro, Microsoft HoloLens, and the built-in capabilities of everyday mobile devices, numerous use-cases for AR have emerged. These range from innovations in futuristic interfaces to globally popular AR games like Pokémon GO by Niantic (see Figure 1) [5][6]. In recent years, AR has been utilized across a variety of industries. For example, in the entertainment industry, it is often used to allow children to render their favorite characters or toys right in their living rooms [7][8]. Retail companies such as IKEA [9], allow customers to visualize furniture in their private homes before making a purchase. In manufacturing, AR helps workers by enhancing assembly processes and allowing them to visualize structures or 3D models in a real-world environment. These examples highlight how some of the different work sectors leverage augmented reality technology for different purposes.

2.1.1 The Reality-Virtuality Continuum

AR as a medium falls into the Reality-Virtuality Continuum originally proposed in 1994 by Milgrim Et al. [10] and later revisited by Skarbez Et al. [11] in 2021.

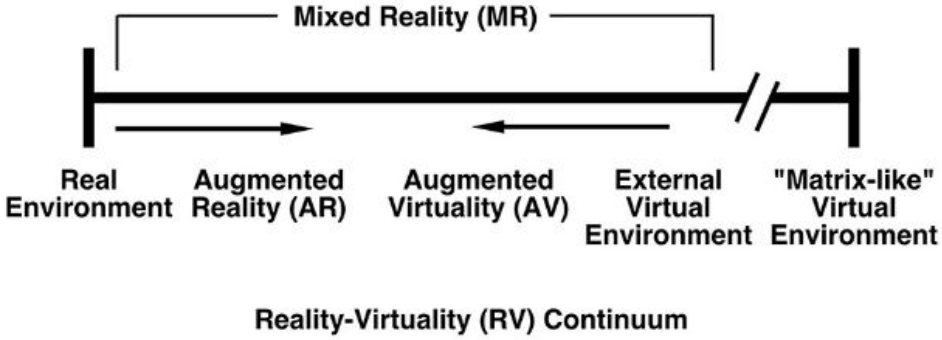


Figure 2: The revisited Reality-Virtuality Continuum, with the added Matrix-like virtual environment extending from the original Mixed Reality Continuum [11].

Understanding this continuum provides valuable insight into where AR falls into the spectrum of Mixed Reality (MR), ranging from a real environment to a virtual environment. Understanding AR's relation to this spectrum means you can compare and have considerations on the other MR Mediums such as Virtual Reality (VR), Augmented Virtuality (AV), and even the Matrix-like virtual environment simulation. Choosing AR as the main direction as opposed to something like AV or VR, which can have a much bigger potential for immersion and interaction with the built-in capabilities of real-world object manipulation and specialized headwear and controllers, comes

down to AR having a much easier barrier of entry for people to use. AR does not necessarily require external specialized equipment, such as headwear or other controllers, and usually runs on a mobile or tablet device. This means that there are many more restrictions, not only in terms of the available interactive elements being limited to the device's touch screen, gyroscope, microphone, geolocation, etc. But also the hardware specifications of each device. As an AR application could be built to support most devices, new and old, some older models of Android or Apple devices might not have the same capabilities with camera resolution, battery life, or general rendering and processing performance. These limitations and advantages will be worth considering during the design process in future development of the AR application.

It is worth noting that AR can also involve specialized equipment such as the Apple Vision Pro or Microsoft HoloLens 2 headwear devices, which are more commonly categorized Augmented Reality/Mixed Reality devices (see Figure 2). We will exclusively focus on the direction of AR for mobile and tablet devices, as we value the universal accessibility, without the need for specialized equipment as one of the primary advantages.

2.1.2 Technical Review of AR

In terms of a more technical angle of AR in a development sense, we also need to consider some of the software development kits and libraries.

ARCore and ARKit are the two primary state-of-the-art SDK's for building AR applications for Android and iOS respectively. These SDK's allow developers to easily integrate essential AR functionality so that custom applications can use the device locomotion, camera, geolocation, and even LIDAR, to best as possible create a seamless integration of digital content into a real-world environment. ARCore and ARKit have different advantages and downsides, mostly dependent on their device compatibility, as Android and iOS devices have different hardware and operating systems. Minor performance differences between the two are not much of a concern and will not affect our decision when choosing the device for study. There are features however, such as LiDAR (Light Detection and Ranging): a method similar to Radar using light as a remote sensor [12]that are much more compatible on iOS devices with ARKit. This could provide a very noticeable performance improvement when it comes to the Simultaneous Localization and Mapping (SLAM) technique used for mapping the environment digitally. LiDAR is not exclusive to only iOS devices, as some Android devices also support it, however, it would be optimal to develop for iOS devices with ARKit as this would give us the highest probability of using LiDAR, even on older devices. While a fully developed application optimally would support both iOS and Android platforms, as they both share a large part of the device market, the scope of this project aims to prioritize the development of a functional prototype to test the upcoming FPS as opposed to releasing it to the public (see Section 2.4).

There are many ways to develop an AR Application, from using external development platforms for AR or WebAR such as 8thwall [13], to building a computer vision project from scratch using

OpenCV. However, due to our previous experience [14][15] with the Unity3D game engine [16], we instead opted to search for solutions that support the development of AR applications within Unity. While one of these solutions; Vuforia [17], seems optimal, being compatible with Unity and much more, ARFoundation [18] has great integration in Unity specifically. As both of these solutions are based on using ARKit and ARCore, the efficiency and consistency for tracking are very equal. ARFoundation, however, has many easy-to-implement Unity-compatible features such as depth maps, occlusion, object tracking, environment probes, GPS anchors, and more. ARFoundation is also well supported, and in turn future-proof, has documentation readily available, and is completely free to use. Of course, as we plan to develop primarily for iOS, by using ARFoundation, we also ensure having LiDAR support which could prove to be an essential component [19].

Lastly, the development platforms we will be using are Unity, ARFoundation, and ARKit (for building iOS Applications). Converting and building the application for other platforms, such as Android, Hololens, or Apple Vision Pro should be relatively straightforward without major issues in the development and build pipeline.

While the project *The Hologram in My Hand* [20] by Bach Et al. uses Hololens and physical markers (image targets) which is a different direction than this project, they do focus a lot with 3D data visualization and AR. Specifically, they mention that one of the benefits of using and navigating 3D space in AR is the additional degrees of freedom (DoF). DoF refers to the number of basic ways an object can move through 3D space (see Figure 3).

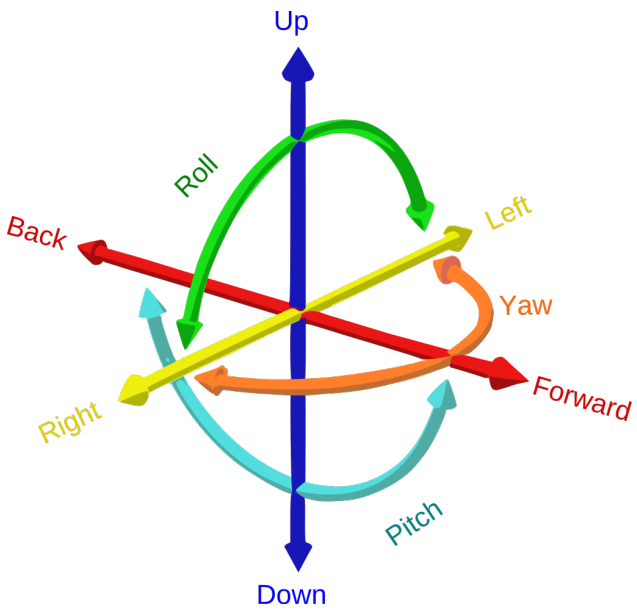


Figure 3: A Gizmo representing the 6 degrees of freedom; Translation: forward/back, up/down, left/right & Rotation: yaw, pitch, roll [21].

With AR as a medium, navigating around 3D space using a 6 DoF camera perspective through

the device locomotion could enable a much easier and intuitive controlling mechanism for visualization in 3D space. This in addition to the touch screen canvas provides even more dimensions of interaction and creates the basis of our initial problem statement (see Section 1.1) that AR as a medium can provide a better user experience than a more standardized desktop application that relies on a mouse, keyboard, and monitor for navigating and visualizing 3D environments.

2.2 Data Visualization

The graphical representation of data can be traced back centuries when maps and diagrams were used for navigation and understanding the world around them. During the Renaissance, it slowly evolved into the complex graphs that we know from today and it allowed us to view data in a multifaceted way. This shows that humans have always sought a way to simplify and communicate complex information graphically [22].

Data visualization is the process of displaying data in a visual/graphical context to describe its characteristics and significance. Data visualization is very good for simplifying complex data that would otherwise be uninterpretable and allows you to analyze data and gain new insights that would otherwise not be possible. There are many varying definitions of what data visualization means depending on who you ask. In the book *The Visual Imperative* by Lindy Ryan [23] the definition of data visualization is stated as:

> "Thus, we can define data visualization as a visual display of information that is transformed by the influence of purposeful design decisions with the intent of encoding and conveying information that would otherwise be either difficult to understand or unlikely (or impossible) to connect with in a meaningful way." [23].

The impact of data visualizations cannot be overstated in today's society, as it allows for the transformation of complex datasets into something visually understandable and feasible. Studies have shown that the visual representation of information, like through charts or images, leads to better memory retention compared to when the same information is conveyed via a text or audio format [24]. It is used in a wide range of business and scientific disciplines. By converting numerical data into accessible visual representations, such as graphs and charts, otherwise complex data becomes more easily accessible. It furthermore enhances the decision-making ability of anyone using it, as decisions can be made based on factual data and not assumptions [22].

Most modern companies utilize business intelligence tools such as Microsoft's PowerBI or Tableau to analyze trends and support data-driven decision-making. In these tools, you connect to a data source and set up various data visualizations that are finally displayed on a dashboard. Additionally, the emergence of web technology such as JavaScript and Python's Jupyter Notebooks has also made it easier to create visually appealing, interactable, and highly customizable plots and graphs with minimal experience required. Some of most popular and well-known data visualiza-

tions today include line charts, bar charts, pie charts, box plots, and scatter plots [25].

2.2.1 2D Data Visualization Types

Understanding the usefulness and what a specific data visualization provides is an aspect that we will delve more into. Examining the various plots, charts, and maps helps to paint the picture of the most essential knowledge that we should aim our focus toward. The first chart we want to examine is the *line chart*.

Line charts are essential for visualizing data trends over time, offering a clear perspective on how variables evolve. According to Tufte's principles in *The Visual Display of Quantitative Information* [26], line charts allow for the identification of patterns, trends, and changes over a period of time, making them invaluable in financial analysis, weather forecasts, and any area where temporal changes are critical. As mentioned by Tufte, the line chart should be simple and provide clarity. Nonessential elements such as grid lines or labels should be minimized to direct the viewers attention towards the data's evolution over time. Data ranges and scales must be selected carefully to represent without distortion, as they can heavily affect the viewers interpretation of the data [26]. While line charts are good for showing trends over time, bar charts are better for for comparing quantities across different categories.

Stephen Few's *Show Me the Numbers* [27] suggests that bar charts are highly efficient at displaying differences in magnitudes making them ideal for comparisons, such as sales performance across different regions or demographic data. Essentially, bar charts encode quantitative values as length (height) along a single vertical or horizontal dimension. This allows the viewer to simply compare the lengths of the bars making it a very simple and accessible type of graph. Bar charts can encode multiple sets of values by also utilizing the width or the hue/color of the bars. However, this can be less intuitive for the viewer to interpret but can be useful in some scenarios. Bars can also be stacked on top of each other to display part-to-a-whole relationships, and to display the amount each part contributes to the whole [27]. Like bar charts, pie charts are also used to illustrate parts of a whole, but are more usually used for situations with a limited number of categories.

Pie charts are used to illustrate parts of a whole in scenarios with a single or limited number of categories, offering accessible insights into distribution and shares. They are known for their simplicity, with each slice representing a proportion of the total. However, their efficiency diminishes with complex datasets or numerous categories, where distinguishing between slices becomes a problem. As known, both pie and bar charts can represent parts-to-whole relationships, bar charts, particularly when unstacked, provide clearer comparisons between segments. The choice between pie and bar charts depends on the complexity of the dataset and the intended clarity of representation. For simple, part-to-whole ratios, pie charts are suitable. For detailed comparisons and with multiple categories or complex relationships, bar charts are preferable. Nevertheless, the selection should simply aim to enhance the viewer's data interpretation and comprehension [27].

Box plots are an efficient method for analyzing and summarizing data distributions, offering insights into the spread and central tendency of the data. It is particularly useful for identifying outliers. It was developed by John Tukey, a statistician known for his many contributions to the visual presentation of quantitative data. Box plots encode a range of values within a box, denoting the distribution's quartiles. The plot includes a middle line marking the median and whiskers extending from both ends of the box to represent the full extent and spread of the data [27]. While pie and bar charts are useful for visualizing categorical data, box plots can be used to analyze numerical data distributions. Another type of data visualization is scatter plots which focus on examining the relationship between two variables.

Scatter plots are a key instrument for analyzing the relationship between two variables and can reveal correlations, trends, and clusters in a given dataset. It essentially plots two sets of quantitative values against each other, one along the x-axis and the other along the y-axis. It allows for insights into how the change of one variable affects the other variable. The interpreter can then decide whether the relationship is strong/weak or positive/negative by observing the resulting pattern on the scatter plot. By visualizing data points as dots and possibly a line of best fit or trend line, scatter plots simplify the complex relationship between variables. Scatter plots are less accessible than some of the graph types that were previously presented and are mostly used by statisticians or data analysts. The same is true for box plots and as a result, they are used less by organizations and companies when visualizing their data [27].

The data visualizations presented in this section are most commonly found in 2D solutions. Their use concerning an AR application must be explored further, as it can be limited. As a result, we must explore the 3D data visualization options, aiming to find a solution for the initial problem statement (see Section 1.1).

2.2.2 3D Data Visualization Types

Before delving into 3D data visualizations, it is important to mention some of the limitations of 2D data visualizations. The primary limitation of 2D visualizations is their inability to depict depth and variable relationships for datasets with more than two dimensions. This inadequacy usually results in having to use multiple visualizations/graphs in order to convey all dimensions of the data, which complicates the data interpretation and insight. The transition to 3D visualization addresses these issues by using the spatial dimensions to represent multidimensional or complex data. It furthermore allows for real-time data interactions due to the static nature of 2D data visualizations. This is especially useful in areas like geographical data, city planning, or any field where seeing the big picture and the tiny details at the same time is important. With 3D data visualizations, you can usually interact with the data in real time, turning it around and looking at it from different angles, which is not possible with flat charts. 3D visualizations potentially require a higher level of user interaction and cognitive engagement, especially in an augmented reality environment. They can be explored from any angle and this can affect the interpretation

of the visualization. Furthermore, it can also cause complexities in regards to user navigation. It is therefore important to consider the balance between enhanced immersion and cognitive load which will be explored in Chapter 5.

Advances in 3D computer graphics and web technology in recent years have brought new tools that allow you to turn complex data into dynamic three-dimensional visuals, with less code [28]. The rest of this chapter will discuss the various technologies used for 3D data visualization in the modern world. 3D data visualization shows the dimensions of height, width, and depth. Creating a 3D model typically involves the following steps: scene, geometry, and final rendering. A modern example of 3D data visualization could be a 3D scatterplot (see Figure 4. 3D data points/symbols are scattered on three axes to explore the relationship between three variables. Additional dimensions can be added by using the shape and color of data points/symbols The usefulness of 3D data visualization have generally been questioned many times [29]. For example, if one tries to find the X, Y, Z coordinates of a data point it is immediately obvious that this task is nearly impossible, highlighting the inability of the human eye to process 3D images compared to 2D images. Furthermore, a common issue with 3D scatter plots is the occlusion of data points. To unlock the potential of a 3D scatter plot, interactive features are a necessity, such as letting users change the viewing angle and zoom into details.

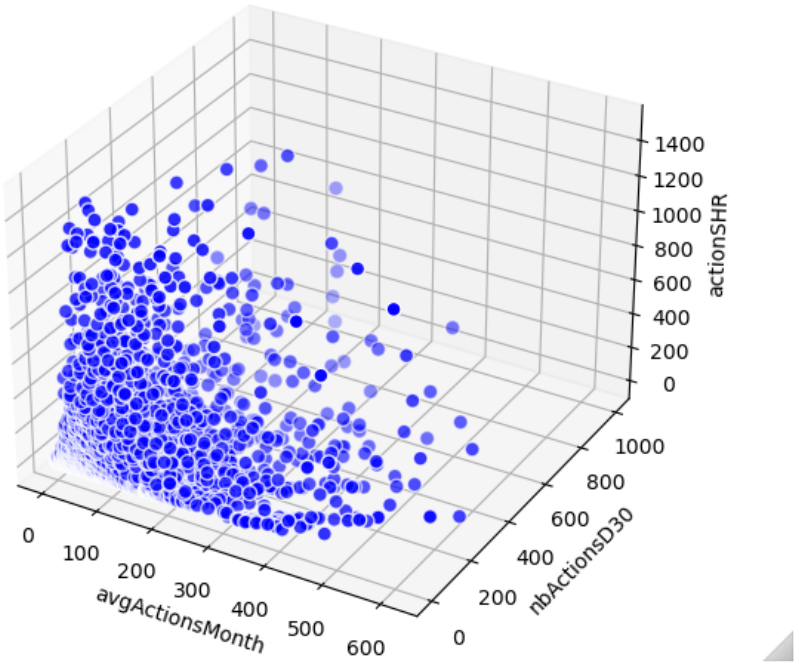


Figure 4: A 3D scatter plot example, utilizing all three dimensions by occupying the X, Y, and Z-axes.

Another type of 3D data visualization is the surface plot. Surface plots convert data points into

real 3D shapes, improving viewer interpretation compared to 3D scatter plots (see Figure 5) but requiring that each XY coordinate must be associated with a numerical value, which usually represents height or depth. This basically allows us to understand the relationship between the two independent variables and the dependent variable. Landscape maps is one particular use-case for this, where you could easily use the dependent variable to display a variables such as elevation or any other kind of numerical value. They are widely used in many scientific fields e.g., topographic imaging. Surface maps can be enhanced by incorporating visual cues into AR such as the use of textures or animations to represent variables such as wind speed across a surface.

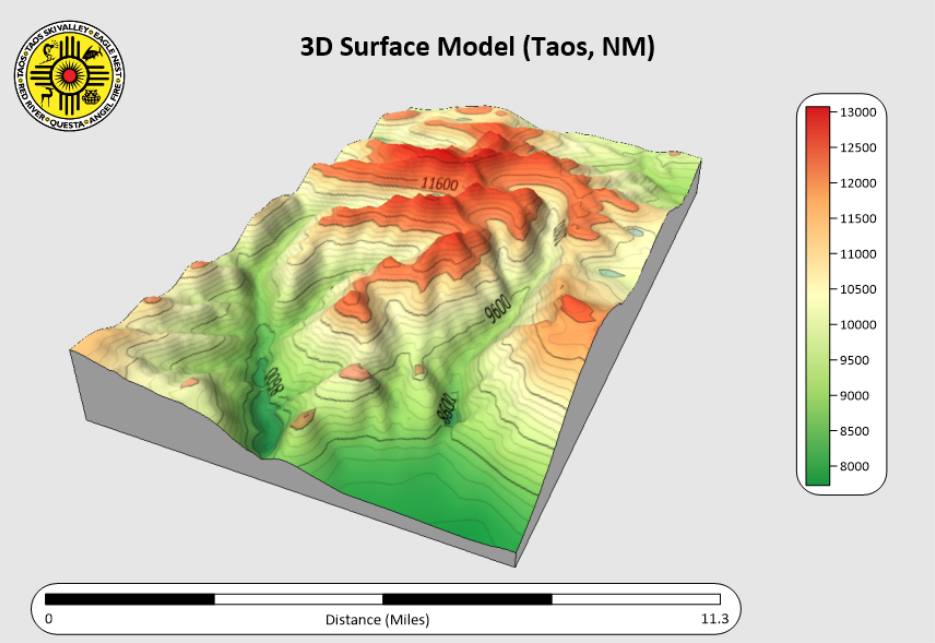


Figure 5: A surface map created in Surfer, a software produced by Golden Software that helps engineers interpret complex geospatial datasets and transform them into understandable models.

Some of the visualizations described in the 2.2.1 section are better suited to a 3D AR environment than others. Pie charts and bar graphs in particular comes into question. The immersive nature of the AR environment allows the user to explore data from any viewing angle or location, which can distort and change the way data is interpreted/perceived. For example, the slices of a pie chart or the relative height of bars in a histogram may be skewed and difficult to accurately ascertain as the user moves around the visualization. Therefore, it might be better to use other visualization techniques that focus on relationships and trends as opposed to visualization techniques that portray differences in size or scale, such as pie charts and bar charts. However, line plots would be more suitable for 3D data visualizations compared to bar or pie charts. By adding an extra dimension, line plots can be used to portray more complex datasets, including features such as time, depth, or any other quantitative measure. This could reveal patterns and trends that are not visible in traditional 2D representations. In physics, 3D line charts could for example be used to show the

trajectory of objects through 3D space.

After describing some of the most commonly used 3D visualizations, the issue of interpretability remains at large. The issue of perspective distortion must be considered, as it changes the perceived scale of data depending on the viewing angle. The farther away an object is the smaller it appears, which is somewhat problematic when comparing the size or length of objects. This can usually be circumvented by using an orthographic view with no perspective, but in context of AR this would not be possible. Occlusion is also a substantial problem to consider when moving freely around a 3D space.

2.2.3 3D Data Visualization Technologies

The following chapter will explore the field of 3D data visualization technologies and will explore the current tools that permit users to interact, visualize and analyze data in a three dimensional space. The main focus will be on solutions that support the 3D visualization of data in any given context, to explore and gain insights into the current landscape and state of the art technologies used for 3D data visualization. By investigating well-known platforms like PowerBI, Tableau, Matplotlib, and Plotly we want to better understand how these tools and applications support the 3D visualization of data.

PowerBI is the leading business intelligence tool and is developed by Microsoft [30]. It is widely recognized for its extensive data visualization capabilities. Although PowerBI does not naively support 3D visualizations, they have addressed this limitation, by allowing for the integration of Python scripts directly within in the PowerBI framework. This essentially allows you to take advantage of powerful data science libraries such as Numpy, Pandas and Matplotlib [31, 32, 33]. By incorporating Python scripts, users can take advantage of Matplotlib's extensive features to generate interactive 3D plots and charts directly within the PowerBI environment. The flexibility of Python means that almost any type of 3D plot can be rendered inside PowerBI dashboards as long as it is supported by Matplotlib. This approach not only adds more visualization features to PowerBI but also offers the creator a high degree of customization in how the data is presented as they can tweak the Python code to adjust almost every aspect of a visualization. Another well-known business intelligence tool is Tableau, which is developed by Salesforce [34]. Unlike PowerBI, Tableau does not support 3D data visualizations as, according to many data scientists and analysts, goes against best practices regarding the visualization of data. This argument stems from the idea that 3D visualizations can often lead to misinterpretations of data due to issues with perspective, occlusion, and over-plotting, which can obscure interpretability [29].

Matplotlib is a widely popular Python library for data visualization that is known for its ability to create high quality graph, charts and figures that are highly customizable and can be presented in both static, interactive and animated formats [33]. It was created by John D. Hunter in 2003

and has since become the foundation of numerous other Python data visualization libraries that are built on top of Matplotlib base, such as Seaborn for statistical and visually appealing plots, or Plotly for interactive plots [35, 36]. Matplotlib works seamlessly with many of the most popular data manipulation libraries in Python such as Numpy, Pandas or Polars [31, 32]. Matplotlib can be extended with the mplot3d toolkit, which allows for the creation of wide variety of 3D plots and charts, such as 3D line plots, scatter plots, surface plots, contour plots, etc. While Matplotlib's 3D capabilities are extensive and powerful, they come with some limitations. For instance, rendering large 3D datasets is very resource intensive, and the interactivity of the plots can be limited.

D3.js is an open source JavaScript library for creating interactive 3D data visualizations in the web browser [37]. It uses web technologies like HTML, CSS, and SVG files to efficiently visualize and let users interact with data in a 3D space. It essentially provides a framework for combining complex data sources into a interactive visual context, since each object is essentially pure HTML, CSS, and SVG and can be manipulated with the browser's Document Object Model (DOM). For each data point, D3.js creates a visual object, such as a circle or a rectangle. This is a different way of visualizing data, usually done through a pre-built charts API. Many popular 3D visualization libraries are built on top of D3.js for these exact reasons, like Plotly.

Plotly is another Python library that can be used for creating 3D visualizations. Plotly is an powerful open-source library well known for its visually appealing and highly interactive charts [36]. Built on top of D3.js and stack.gl, it basically allows dynamic and interactive 3D visualizations to be rendered directly within web browsers with only minimal code needed. The tool is flexible as well since it supports various programming languages like Python, R or JavaScript. With Plotly one can make different types of diagrams but in this study, we will focus only on 3D visualizations. For instance you may use 3D scatter plots, 3D lines charts and many more. These charts look modern and attractive and at the same time being interactive, allowing the users to zoom in/out, change viewing angle or highlight/slice the specific data you are interested in.

2.2.4 AR Data Visualization Technologies

Having explored the capabilities of current 3D visualization tools, we will explore how augmented reality can support the visualization of three dimensional data. This section will primarily focus on AR data visualization technologies grounded in recent studies and developments in the field. We aim to better understand AR's capacity to enhance data visualization processes by providing more immersive and interactive experiences that might make complex data more engaging.

VisualLive is an AR tool by Unity that allows you to bring 3D models and visualizations into the real world [38]. It allows users to visualize complex designs on-site, supporting the import of over 70+ of the most common 3D file formats (Autodesk Revit, Navisworks, etc.). VisualLive is a somewhat unique AR tool, as it supports both mixed-reality and augmented reality devices. It currently supports the HoloLens, iOS, and Android. As can be seen in Figure 6, this workflow

of VisualLive is very versatile and supports a lot of different imports and outputs. The tool was primarily developed for computer-aided design (CAD) applications and the construction industry, specifically building information modeling (BIM). VisualLive essentially allows for the interaction with full-scale 3D models in a real-world environment. Construction sites would normally be fitted with QR codes, which allowed the workers to scan them and visualize future structures in the context of their intended location. As of 2024, Unity has stopped the future support and maintenance of the VisualLive AR solution to focus on other projects.

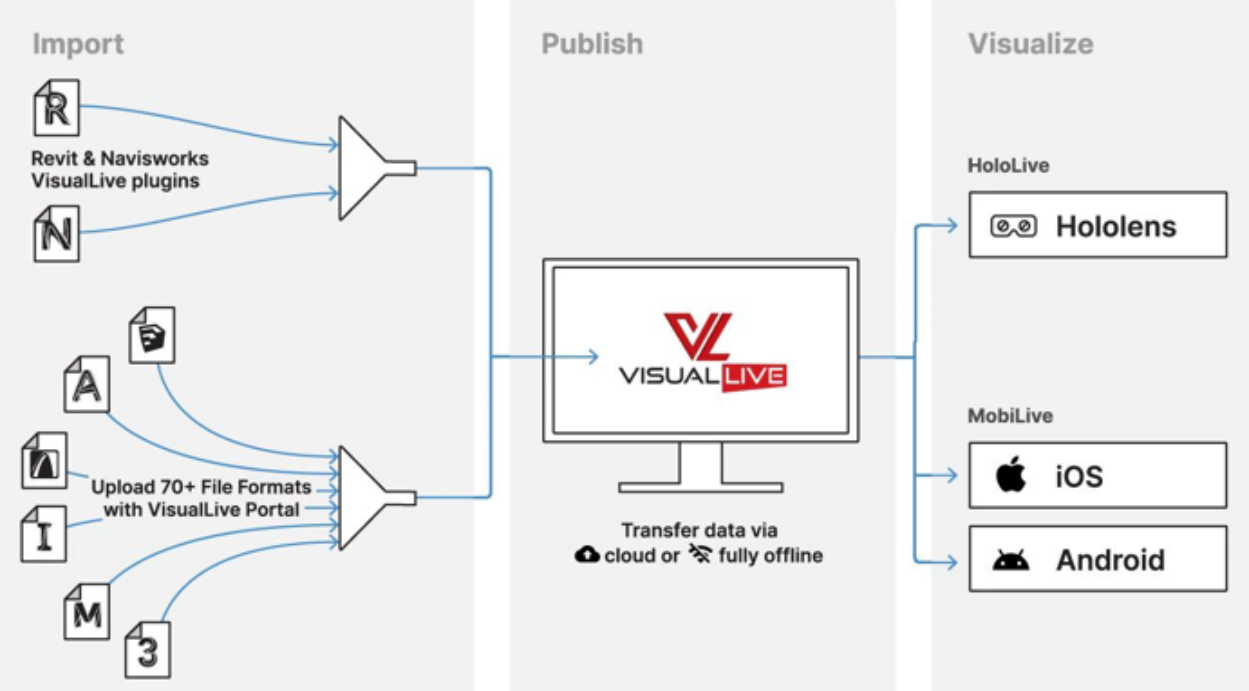


Figure 6: Workflow diagram for Unity's VisualLive AR: Import BIM/CAD files and publish data to the cloud or for offline use, and visualize designs on-site with HoloLens for HoloLens and MobiLive for iOS and Android devices [38].

VisualLive has since evolved into Unity Industry. Unity Industry allows sectors like automotive, manufacturing, and architecture to create custom 3D experiences. These experiences can be for AR, VR, mobile, desktop, and web platforms, making them versatile for various applications. Unity Industry allows users to transform their CAD and 3D data into immersive Unity experiences and apps, alongside Unity Mars for augmented reality experiences [39, 40]. The transition from VisualLive to Unity Industry underlines their commitment to being in the forefront of real-time 3D application development, by making the tool available for a broader spectrum of industries and making it extremely versatile in terms of output. Unity Mars allows for the efficient creation of AR experiences by integrating easy-to-use and versatile templates and a robust testing environment within the Unity Editor, allowing for a nicely streamlined development workflow when developing for AR. It furthermore supports cross-platform development for all their supported devices (HoloLens, iOS, and Android).

In the study "MARVIS: Combining Mobile Devices and Augmented Reality for Visual Data Analysis" by Langner Et al, the concept of combining mobile devices with augmented reality for visual data analysis is explored [41]. The "MARVIS" framework merges mobile technology with augmented reality for enhanced data visualization. This approach offers new interaction and visualization methods hoping to expand the capabilities of mobile devices for data analysis and visualization. Their approach utilizes augmented reality by using a mixed-reality headset to complement data visualizations shown on mobile devices, improving the user's ability to analyze the data by providing spatially augmented visual cues. They are addressing some of the difficulties when using mobile devices for data visualization, specifically through the use of AR Head Mounted Displays, which are implemented as an addition to the 2D charts shown on the mobile devices. This essentially allows for the display of extra dimensions or attributes, which would otherwise not be possible (see Figure 7).

Figure 7: This image is from the perspective of the MARVIS user. A 2D chart is displayed on a mobile device with 3D visualizations displayed above it, achieved by using the Microsoft HoloLens [41].

Eliminating the need for secondary devices, such as AR Head Mounted Displays (HMDs), ensures a more accessible experience. The MARVIS framework furthermore explores the collaborative potential of AR with multiple devices. This would allow multiple users to share and interact with data in real-time. The AR part of the solution was developed in the Unity game engine while the mobile device client was built as a web application in JavaScript, using libraries such as D3.js for visualizations. The study furthermore included a large exploration of user interaction mechanisms within an AR environment centered around data analysis and visualization. It describes how users can manipulate data visualization through intuitive gestures and movement. Some of these data interactions will be explored in the following section 2.2.4.1.

In the study "The Hologram in My Hand: How Effective is Interactive Exploration of 3D Visualizations in Immersive Tangible Augmented Reality?" by Bach et al, a controlled user study was conducted to compare three visualization environments for exploring common 3D tasks [20]. The three visualization environments which were all based on augmented reality technologies included an AR Head Mounted Display (Microsoft HoloLens), a handheld tablet and a traditional computer desktop setup. Their research was aimed at understanding how the various environments affect the human perception and interaction capabilities but also how they impact the

understanding of 3D data visualisations. The study found that each environment had specific strengths for specific tasks, but that the desktop environment was the fastest and most precise for a large majority of the tasks. This study will provide valuable insights for our project, although the solution implementation might differ somewhat. It is important to note that study found handheld tablets offered a more intuitive and natural interaction for 3D visualizations compared to desktop environments [20]. They furthermore highlighted the advantages of handheld devices by pointing out their ease-of-use and portability, but also suggested that they lack far behind in speed and precision compared to desktop environments. They also suggested the potential for increased cognitive load in AR environments due to badly designed interaction. Although they concluded desktop environments were superior, they highlighted the importance of context in choosing the right visualization environment, suggesting that AR environments functions optimally when immersion and spatial interaction add significant value. The decision to pursue an AR environment, despite desktop environments efficiency, is motivated by the unique advantages that AR can offer in specific contexts. Our project seeks to take advantage of these strengths by developing a solution where AR's immersive and interactive capabilities are utilized to enhance data visualization, beyond what is capable with desktop environments.

2.2.4.1 Data Interactions in Augmented Reality

In this section 2.2.4.1, we will delve into some of the options that are available designing interactions in an augmented reality environment. The main objective will be to find user-friendly interactions that minimizes cognitive load and enhance the overall user experience.

Some of the interactive that are commonly available in data visualizations tools, are UI-based widgets such as sliders that allow for features such as zooming, rotating, highlighting or cropping and are typically placed beneath or besides the chart itself. These could very easily be replaced by touch-based gestures in a mobile based AR environment. These common touch-based gestures are very easy to learn and most people are already familiar with common touch-based gestures and interactions and it will therefore likely offer high precision and familiarity to users, although UI-based widgets could still be useful in some scenarios. This would allow the user to interact with the data by touching data points, axes or by zooming in on specific parts of the data.. When using tablets or mobile devices there could be some issues with tap and drag gestures due to the relatively small screen size of some mobile devices and tablets.

Besides the more common tap-based gestures and interactions in augmented reality (AR) environments for mobile based devices, it is important to explore the possibilities of spatial interactions. Spatial interactions is essentially physical movement in 3D space and allows the user to navigate and view data from different angles and locations, and basically allows the user to crop data by utilizing the spatial dimension. This removes the need for a zoom and panning interaction that is commonly found in many interactive data visualization. It forces the user to move around in space but may not be suitable for all users or contexts, especially where real physical space is lim-

ited. Therefore, utilizing spatial interaction could likely improve the interaction and immersion of 3D data visualization.

2.3 User Experience

As we try to answer our initial problem statement (see Section 1) we want to know how to optimize user experience in an augmented reality application. In this case, user experience becomes the key to unlocking a deeper understanding of the users. Gaining knowledge on how to optimize user experience and learning what to specifically focus on, fundamentally strengthens the end-product. Consequently, we consider the cognitive processes important to acknowledge and examine. By exploring human cognitive resources, we seek to identify challenges and navigate the field to help us recognize pitfalls and enable the cognitive processes most optimally.

2.3.1 Target Group

The current objective for this project is to figure out how augmented reality can be used as a medium for enhancing data visualization to optimize user experience (see Section 1.1). Researchers, students, and companies that work with statistics would be the most optimal target group for this project. Researchers would have state-of-the-art knowledge of using statistics for publications, meaning they would be a great fit acting as the sample population on the topic of statistics. Students, and especially last-year students at universities, can provide us with data about what they see as the most important statistics to use, providing a fresh and new perspective on the field of statistics. Companies would emphasize what statistics are being used from a more competitive and economic perspective, more likely to provide information on their company's specific statistical needs. As a result, and wanting to narrow in on what specific target group to use, we have examined the field of work that has comparisons to our project [20][42][43]. More specifically, we examined the three publications target groups as well as their population, and it was found that it is possible to not have a specific target group in order to receive positive results. However, working with a population that is familiar with statistics should be a priority, as they might have a greater understanding of data visualization in general. Furthermore, students from the Technical Faculty of Aalborg University in Copenhagen can be an option as participants as they have knowledge of using technologies and working with statistics. They are easily accessible and will be consistent in their way of evaluating the product, reducing potential outliers. Additionally, they have a fundamental understanding of statistics and data visualization. Having a population that have knowledge within the desired area will in most cases reflect or mirror an authentic working environment and its expertise in augmented reality and data visualization.

As a result, researchers and students at the Technical Faculty of Aalborg University in Copenhagen will act as the target group, providing preliminary information on their usage of statistics. Obtaining this data will provide the project with a starting point and direction on what to improve in the domain of statistics. In later iterations, the target group will consist of individuals who are

familiar with statistics, as the likelihood of them understanding the data visualization application is greater. For this, students with knowledge of using technologies and working with statistics can act as the primary population. This will ensure that the population has relevant knowledge within the field of work, to some degree reflecting a real working environment in this domain. In addition, examining individuals with different levels of statistical knowledge could be the substance of an interesting case. Categorizing individuals into three knowledge levels of statistics can help to label the participants and later act as a supplement to the research question. Moreover, students are chosen as they are more accessible to reach then researchers and the overall population group is bigger. That being stated, if researchers are available for the final testing, they will be allowed to take part.

2.3.2 User-Centered Design

When examining the field of user experience a few names come to the surface. One of those is Donald A. Norman. Norman is known for his work and expertise in user experience. Norman states in his book *The Design of Everyday Things* that: "*The human mind is exquisitely tailored to make sense of the world. Give it the slightest clue and off it goes, providing explanation, rationalization, understanding*." [44]. Furthermore, he advocates for User-Centered Design (UCD) [44]. UCD is an approach that encompasses a few principles and concepts that aim at developing products that prioritize the needs, preferences, and experiences of the users. The work on UCD originated from his laboratory work at the University of California, San Diego in the 1980's, where the term since got widely known and used as a result of the publication of the book; *User-Centered System Design: New Perspective on Human-Computer Interaction* [45], and later *The Psychology of Everyday Things* [46]. He developed his framework on UCD through the years, and it has led to splitting the task of creating and developing a product into four divisions: 1. Understanding the context of use, 2. Specifying a user's requirements, 3. Design a solution, and 4. Evaluate the solution against the specified requirements (see Figure 8) [44].

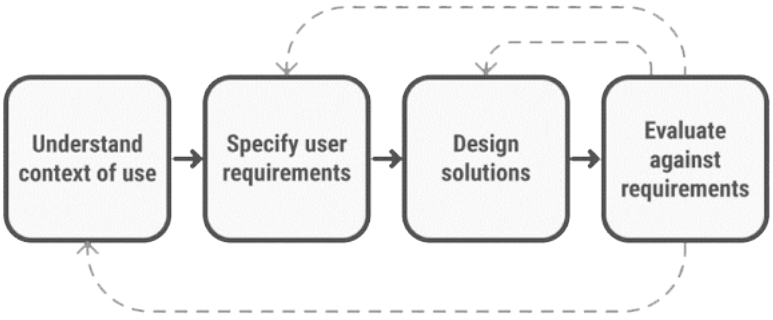


Figure 8: The four phases of the UCD approach: context, requirements, design, evaluate [47].

2.3.2.1 Understanding the Context of Use

Delving further into what the four divisions specifically entail, you discover that it becomes important to involve various factors e.g. the physical environment, social norms, cultural practices, users goals, and much more. In order to gain an insight on this you must understand the context of use. Norman argues that designing without considering the context of use can lead to products that are difficult or frustrating to use, or in general unusable [44]. Gaining this insight before starting the development of the application is vital as considering the exact use case and environment the application will be used in, becomes an important factor in the path toward the end product.

To obtain this information you can conduct *user research*. Deciding early on how the data will be used in a study is vital, and for user experience data you can either use *formative* or *summative* methods. Albert and Tullis [48] present the formative user experience method as the act of periodically checking in on a dish, as a chef. While it is being prepared you can make a few adjustments to impact the end result. Throughout making the dish, the chef periodically evaluates, adjusts, and reevaluates the dish. The same can be said to happen in formative user research. Throughout the creation of the product it periodically gets evaluated, and from one iteration to another you identify the issues, collect recommendations, make adjustments, and then repeat the process, until the product comes as close as possible to being perfect [48]. The goal of formative user research is to gain knowledge on potential issues prior to release. This is supported by the agile iterative approach of product design. Jumping back and forth between design, implementation, and evaluation helps to develop a more polished product with little to no issues due to combating them early in the process. However, formative user research is ideally done before the design is finalized. The earlier the formative evaluation is done, the more impact the evaluation will have on the design. Albert and Tullis recommend a total of 8 to 12 participants as the sample size when using the formative user research method. This sample size is large enough to identify problems, but not great enough to identify preferences. In general, the formative user research method should be able to help with discovering the most significant issues in a product as well as what aspects or features work well for the users and what improvements should be considered for the next iteration [48].

Albert and Tullis continue with their cooking metaphor for explaining the summative user research method. In this example, they state summative user research as evaluating the dish after it comes out of the oven. They see the user experience specialist or researchers as a food critic who evaluates a few dishes and maybe compares the dishes to other restaurants [48]. Essentially, the goal of summative user research is to evaluate a product and compare it to how well it fits the objectives. The summative method focuses on evaluating products against a set of criteria. Usually, summative user research can help answer questions about the overarching UX goals of a product as well as how it compares to other similar products.

In order to gain knowledge and insight into what users specifically want in a product, you can conduct interviews. When working with interviews, it must be considered whether you want to use a structured, unstructured, or semi-structured approach. The structured interview often provides quantitative data. The approach aims at having the exact same order of questions for each participant. This ensures reliability in the aggregated data as well as the option to compare different samples during different survey periods. The unstructured approach is more or less a guided conversation. It is based on observations that will be used to elicit insights into how a population uses a specific topic. That could be how doctors organize and manage patient encounters, or how researchers interact and utilize statistical methods. Lastly, is the semi-structured approach. This approach often acts as the sole data source for a qualitative research project. Semi-structured interviews are scheduled around a set of predetermined open-ended questions with the opportunity of emerging questions based on the dialogue between interviewer and interviewee [49].

An additional approach for obtaining an insight of the users, is to use *contextual inquiry*. This method is a qualitative data gathering and data analysis methodology and it is based on observation and communication with users in their working environment. Based on Raven and Flanders in 1996, the contextual inquiry method is based on three principles: 1. Data gathering must take place in the context of the users work, 2. The data gatherer and the user form a partnership to explore issues together, and 3. The inquiry is based on a focus; that is, it is based on a clearly defined set of concerns, rather than on a list of specific questions (as in a survey) [50]. As for the first principle, it is important to get in action and observe what the users do in their work life. Observing and communicating with the users in their work environment (or in the context of use) helps to gather data that is different from the type of data you obtain through questionnaires or surveys. In most cases, contextual inquiry data becomes more concrete as a result of experiencing the users work environment, their habits, struggles, etc. [50].

We do recognize that there are more options for gaining an insight into the users perspective, however, we will not delve more into the different options for the scope of this project.

2.3.2.2 Specifying User Requirements

The second step in the UCD framework is to specify the user requirements (see Figure 8). Specifying user requirements is a crucial step as it helps to ensure that the product aligns with user needs. Identifying and documenting a user's needs, goals, and objectives with a product is what this step in the procedure entails [44]. Additionally, obtaining data on the characteristics of the target group becomes vital. Having demographic data as well as their preferences and other relevant attributes helps in establishing user requirements. Containing this knowledge will help to tailor the product towards the intended users [44].

As this is the follow-up step to understanding the context of use, some of the knowledge gained

earlier will be used in this step i.e. the interview and/or observational data. The data gained through understanding the context of use can be used to specify the users requirements for the product. Analyzing the interviews and understanding the respondents will help to form a set of requirements that could act as the success criteria for the product. Acquiring this data early on will help to form a more polished, and hopefully, more in-depth application. More on this later in Section 4.

2.3.2.3 Design Solutions

Having obtained all the necessary data about the users, it is time to design a solution based on that data. Although Norman has not listed a specific set of design principles, several key concepts presented in his work can be taken into account as they align with User-Centered Design [44]. Some of the principles and concepts of the UCD approach are: *visibility*, *feedback*, *affordances*, *mapping*, *constraints*, and *consistency*.

In general, these principles and concepts act as a guideline for the design and development of a product, and in most cases, they are pretty straightforward. For example, making sure the relevant elements and functions of a product or system are easily perceivable by users, is one of the most important aspects of design. Users being able to understand the design and interaction with a system is a fundamental, but very important aspect of every product. Furthermore, ensuring the system provides feedback to the user about the results of their actions, will help them to better understand the system they interact with. This will further strengthen the design by being more transparent as well as ensure a more responsive user experience. This must be supported by affording or suggesting the functionality of design elements. Failing to understand the purpose and use of different elements within a system will result in the users not being able to understand and use the system. Users should be able to easily infer the purpose of the system and its elements based on their appearance, i.e. a cross icon suggests the functionality of closing a tap in the internet browser. Using similar icons as the state-of-the-art applications, is a must, as it establishes familiarity. Furthermore, creating a clear, intuitive, and functional user experience in terms of the controls is important. This could be correlated to mapping, as ensuring a clear relationship between controls and their effects must be intuitive for users to manipulate and achieve their goals. Equally important is the replication of this template throughout the system as this ensures the system behavior is consistent, and therefore easily understandable. Maintaining consistency in the design helps to build a mental model of the functionality of the system, reducing cognitive load. This is especially important when developing new and complex systems that users are not familiar with. In order to further limit mental workload of the users, the design must be constrained. Constraints can act as a guide, ensuring that users do not naively navigate the system, but stay within the desired path. Constraining the users toward meaningful interactions will prevent probable errors. Having these principles and concepts in mind while designing a system will help to bring forward the most essential design elements while preventing pitfalls [44].

2.3.2.4 Evaluate Against Requirements

For the final step in the user-centered approach, you must reflect and assess the system or product based on the success criteria established and based on the users requirements. The process of evaluating is vital for a product to succeed and ensuring the design meets the needs and expectations, is what this step is all about. As the design must address the user requirements, the users must be consulted on multiple occasions. In other words, the user-centered design makes use of an iterative design approach (see Figure 8). Evaluating against these requirements helps the continuity of the product, as regular feedback is collected through user testing and later analyzed. Then the procedure of designing a new and improved solution and evaluating it would keep on going until the user requirements are fulfilled. The results of this iterative assessment will in the end ensure that the user needs align with the product, resulting in a positive user experience [44].

2.3.2.5 Measuring User Experience

In the broad field of user experience, measuring a user's understanding and their easiness around a product is fundamental. However, a plurality of measurements has been created and used in countless combinations [48]. It is important to identify which frameworks to focus on as well as how these frameworks can work together. Combining frameworks will provide a more in-depth evaluation, pointing out the most prominent issues in a system. Albert and Tullis present a great amount of the most used methods for user experience in their book *Measuring the User Experience: Collecting, Analyzing, and Presenting UX Metrics* [48]. An additional framework for user experience that is not mentioned by Albert and Tullis is the User Experience Questionnaire (UEQ), a 26 item questionnaire by Laugwitz et al [51]. It is a widely used questionnaire to measure the subjective impressions of users towards the user experience of a product [52]. Besides the 26 items, the questionnaire includes six factors: *Attractiveness*, *Perspicuity*, *Efficiency*, *Dependability*, *Stimulation*, and *Novelty*. Attractiveness is a pure valence dimension [52]. This refers to the basic emotional reaction of liking or disliking something. This could also be stated as positive of negative valence, however, with pure valence no other complex emotions are involved. In general, this can be seen or described as a simple acceptance or rejection gesture/response. The Perspicuity, Efficiency, and Dependability factors are determined to be of pragmatic quality. This involves interactions with a product that relates to the tasks or goals that the user aims for when starting the interaction. In contrast, Stimulation and Novelty are of hedonic quality, as they describe aspects that relate to pleasure or fun while using a product, and not tasks and goals [52]. Utilizing the UEQ framework later on can provide fruitful data and analyzing it will result in an estimate of the augmented reality medium's overall user experience.

To provide additional details on a users user experience, task performance can be measured. One of the most commonly used usability task performance metrics is *Task Success*. Task Success does not need a large elaboration, the term itself is self-explanatory, it is whether or not a user is suc-

cessful in completing a task [48]. To measure this, each task a user is asked to complete, must have clear end state goals. Additionally, correct answers or criteria must be established in order to define their success in the task. In [48] they mention that having users to verbally articulate their answer is one of the most common ways of measuring and obtaining their success. However, this can at times provide arbitrary information or difficulties in interpreting the answers. In such a case, follow-up questions are required to make sure that the answer is interpreted correctly. Other methods consist of having users to utilize digital tools. This provides a more structured approach but in turn a more time-consuming approach [48]. For numerically interpreting this data you must split their answers into correct and incorrect answers. Correct answers get the value 1, while incorrect answers get the value 0. This is also know as *Binary Success* [48]. This is one of the most common measurements of task success and it can be compared to a pass/fail exam, either you pass (1) or you fail (0). This measurement is appropriate to utilize when the success of the product depends on the user completing a task or a set of tasks. Converting the answers into numerical values helps the analysis and provides easier-to-grasp tables [48].

Another measure for user experience is *Time on Task*. This measure used in correlation with a plurality of measures to helps support the efficiency of a product [48]. Usually, the faster you are to complete a task, the better the experience. Of course there are some exceptions i.e. computer games where investigating the virtual environment and following a story, are in some cases of more importance than speed-running the experience. Time on task becomes an important factor to consider, especially if tasks are performed repeatedly by users. This could be the case of a 3D visualization tool. Most likely, the same procedure and features would be used over-and-over again. Due to this, it is vital that we limit the duration of the fundamental features in the application. Improving the time on task, would most likely improve the user experience [48].

2.3.2.6 Measuring Usability

During the development phase, it would make sense to iteratively assess how effectively users can interact with the application and data visualizations. Usability tests would be optimal for this scenario and would be a useful tool for enhancing the overall user experience of the solution. Usability tests are conducted by observing users as they complete predetermined tasks while using the solution, allowing us to gain insights into user behavior, bugs, and potential difficulties encountered during interaction. According to Nielsen [53], usability tests should primarily focus on identifying usability problems, collecting qualitative data, and assessing the participants overall satisfaction with the product, rather than collecting large volumes of statistical data from a large pool of test participants [53]. The format of usability tests can vary but is usually performed in a moderated setting where a facilitator interacts and guides the test participants. Usability tests will allow us to examine how users interact with the solution and the insights gained from these tests will help guide the iterative design process, allowing us to continually refine and enhance the application based on real user feedback. Usability tests are performed on a small amount of

test participants. As previously mentioned, Albert and Tullis recommend a total of 8 to 12 participants when using the formative user research method (see Section 2.3.2.1). The facilitators of the test should ensure that the testing environment is arranged properly in an environment that supports the requirements of AR usage. Ideally, in a closed environment such as a meeting room with a table and minimal obstacles, allowing the users to roam freely. The participants are normally encouraged to think aloud and the tests can be filmed/recorded, if proper consent is secured. The test facilitators will support themselves with a pre-written script. The test can furthermore be supported by having the test participants fill in a short questionnaire, such as the System Usability Scale after completing the predetermined tasks [54]. The System Usability Scale (SUS) is a common measure to use during usability testing, since it provides a structured and quantitative methodology for the iterative design process. It consists of 10 items which result in a composite score that reflects the system's overall usability. By applying the SUS, you can assess the impact of each design iteration on the user experience which helps prioritize improvements [55].

2.3.3 Cognitive Load

In the context of augmented reality cognitive load plays a vital role. Understanding a human's cognitive resources (or architecture [56]) becomes relevant for effectively developing an augmented reality application. In particular, managing cognitive load (or mental workload) effectively can help users maintain focus and avoid cognitive strain when navigating an AR application. Sweller postulates through his Cognitive Load Theory (CLT) that humans have limited cognitive resources available for processing information, and if these resources become overloaded, learning and performance can be impaired [56].

Cognitive load refers to the amount of information an individual's working memory can process. It has been categorized into three types: *intrinsic*, *extraneous*, and *germane* [57]. However, before diving into the specific categories we must provide a more in-depth context of cognitive load. Based on Gaery's and Berch's work on evolutionary educational psychology [58][59], Sweller has divided *information* into two categories: primary and secondary information [60]. He defines primary information as biological knowledge i.e. the knowledge you gain automatically and without conscious effort - even if the information is voluminous. This is due to the primary information being processed and acquired for many generations leading to learners not having to learn how to process this information, as well as not having to learn how to acquire and store it. Essentially, this is non-teachable as this primary knowledge is acquired automatically early in life. In contrast, secondary knowledge is acquired through teaching. Obtaining this knowledge requires the learner to put up conscious effort as the human's secondary knowledge system requires explicit guidance. Acquisition of these skills becomes vital as their absence will affect a learner's everyday life [60]. Consequently, the AR application developed in this study should try to limit the use of over-complicated information (or secondary information), as primarily focusing on simple-to-understand information and already obtained knowledge, would result in a more user-friendly and intuitive experience.

Now knowing that we, human beings, have two ways of learning and processing information, investigating what the augmented reality application should explicitly accommodate, should be of focus. The three types of cognitive load: *intrinsic*, *extraneous*, and *germane* will be examined in the context of the work of Sweller in his article *Element Interactivity and Intrinsic, Extraneous, and Germane Cognitive Load* [57]. Firstly, we will touch upon the subject of intrinsic cognitive load.

2.3.3.1 Intrinsic Cognitive Load

The term *intrinsic cognitive load* refers to the inherent difficulty in processing information, which is further related to the instructional issues of how this information is presented [57]. Sweller states that: "*The level of intrinsic cognitive load for a particular task and knowledge level is assumed to be determined by the level of element interactivity*". Sweller further states element interactivity as: "*(...) anything that needs to be or has been learned*" [57]. He categorizes this element interactivity into *low* and *high* interactivity levels. Low element interactivity allows for learners to learn new elements with minimal reference to other elements, imposing a low working memory load. In contrast, high element interactivity material requires learners to consider multiple associated elements at once, resulting in a heavier working memory load [57]. However, distinguishing between the concepts of intrinsic cognitive load and element interactivity becomes important. Intrinsic cognitive load emphasizes an important distinction between instructional implications of a task and those based solely on task difficulty, related to a total number of elements involved [57]. Put in simpler terms, designing instructions should not only be influenced by how many elements are involved in a task, but also by how those elements interact with each other. Considering this in the design process should combat the complexity of these interactions and ultimately result in effectively managing the intrinsic cognitive load and supporting users in learning the application without experiencing cognitive strain. An example of low intrinsic cognitive load can be referenced to the experience of learning to ride a bicycle on a flat surface with no distractions. The task itself is quite simple and requires little to no mental effort. It is a well-defined task that does not involve nor provoke any complex cognitive processes, as the focus is only on pedaling, steering, and balancing, resulting in no additional cognitive demands. This low cognitive demand is further helped by the low amount of element interactivity, as it requires minimal reference to other elements, to be able to learn how to ride a bicycle. In contrast, learning to play a complex piece of music on the piano will in most cases, result in a high intrinsic cognitive load. The task itself is complex and requires a significant amount of mental effort. Playing the piano you must coordinate multiple cognitive processes at once, as you must adjust your hand movement appropriately while maintaining tempo and rhythm. Furthermore, the great amount of element interactivity does not help the case of learning to play a complex piece of music. One thing is to remember specific sequences in the music, but to store that information while also having to read, interpret, and act on said information, the element interactivity skyrockets, resulting in a much heavier load on the working memory.

2.3.3.2 Extraneous Cognitive Load

The complexity of intrinsic cognitive load is not the only procedure that affects the working memory. The working memory is also imposed by sub-optimal instructional procedures, and this is referred to as: *extraneous cognitive load* [57]. This greatly involves factors such as irrelevant information, confusing instructions, or poorly structured learning material. Keeping cognitive demands low optimizes the allocation of cognitive resources toward acquiring learning objectives more effectively. Reducing extraneous cognitive load would enhance learning outcomes and show elevated efficiency in users learning processes [57]. Now, why is this important in the case of an augmented reality application? Considering the aspect of developing an AR application for 3D data visualization is a relatively modern direction in technology, streamlining and eliminating unnecessary cognitive demands could result in an effective learning experience. As this direction in technology is somewhat niche, ensuring optimization of the cognitive resources becomes vital as excessive extraneous cognitive load can deplete the cognitive resources quickly, leaving few resources available for learning acquisition. A more in-depth example of low extraneous cognitive load can be related to learning mathematics via a clearly explained textbook. As the instructional material is easy to understand and the textbook provides clear explanations of the mathematical concepts, it will result in a low extraneous cognitive load. Having information logically and uncomplicated presented provides the user with the ability to free up cognitive resources and focus on understanding the concepts more effectively. In contrast, high extraneous cognitive load can be experienced by using a new complex piece of software with a cluttered user interface while providing little to no guidance. Experiencing an environment filled with distractions becomes an obstacle to the learning process. Being overwhelmed with a cluttered user interface and unclear instructions requires the users to spend extra time interpreting the layout along with how the software operates. Putting up additional cognitive effort to understand mechanisms that in reality should be easy to understand, detracts from the ability to focus on the otherwise smart functionality of the software.

2.3.3.3 Germane Cognitive Load

Germane cognitive load refers to the cognitive effort in learning and understanding. It focuses on the cognitive resources allocated to process and integrate information in a way that promotes deeper understanding. If a user experiences a difficult task (high intrinsic cognitive load), and the information is presented clearly (low extraneous cognitive load), the user will use their cognitive/mental resources to understand the task, which will output a greater learning acquisition due to higher germane cognitive load levels. However, if the task information is poorly presented (high extraneous cognitive load), the user will spend their cognitive/mental resources trying to understand the assignment, which in turn will take away their focus of the main task, which is to solve the specific assignment, resulting in lowering the germane cognitive load and learning [57]. The distribution of germane cognitive load is the functionality of devoting working mem-

ory resources toward what counts. Your memory can be seen as a library filled with knowledge. Germane cognitive load acts as an assistant who sorts new material and integrates it with your existing primary and secondary knowledge (see Section 2.3.3). It helps you see patterns, make connections, and build a cohesive mental library. In the case of the augmented reality application, trying to allocate the users mental resources towards the intrinsic material would be of more importance than the extraneous material. The more focus a user puts into the intrinsic department, and with the extraneous department not interfering, the greater the chance they have of obtaining and storing new knowledge. However, it must be noted that this formulation assumes high motivation as well as all working memory resources being devoted to dealing with intrinsic and extraneous cognitive load [57].

2.3.4 Measuring Cognitive Load and Task Performance

To figure out the augmented reality application's influence on peoples cognitive load, we must measure their cognitive load levels. This can be done by subjective, behavioral, and/or physiological measures. In addition, task performance measures can be used to further support the findings. Using a combination of multiple measures (a multimodal approach [61]) can provide an in-depth analysis of a user's experience. Consequently, this section will delve into with some of the most commonly used measures for cognitive load and task performance, in which will be used to determine this project's experimental design.

2.3.4.1 Subjective Measures

Having discussed the interplay between the three types of cognitive load, and the importance of incorporating these learning's into the AR application, we must consider how to measure cognitive load. Sweller, Ayres, and Kalyuga account for different measurements of cognitive load in the book *Cognitive Load Theory* [62]. For this study, finding and using a suitable measurement of cognitive load becomes important given that considering and measuring a user's cognitive load in a task-based approach, plays a significant role in developing the application.

A great number of experiments examining how to measure cognitive load have been conducted through the years. In the early exploration of cognitive load theory, cognitive load was measured indirectly. This is known as an *indirect measure* of cognitive load. The results showed that indirectly measuring cognitive load by measuring error rates and learning times provided sub-optimal results and in turn, stamped the solutions as a negligible indicator of cognitive load. This paved the way for examining different ways of measuring cognitive load, and it was found that *subjective measures* of cognitive load could provide a more telling index [62][63].

Exploring metrics for measuring cognitive load has led us to the *NASA Task Load Index* (NASA-TLX). However, it must be noted that there have been numerous attempts to individually measure the three types of cognitive load (intrinsic, extraneous, germane) some with more success than oth-

ers, due to the impractical nature of subjective measures and the multifaceted nature of cognitive load [61][64][65][66][67].

Nasa-TLX provides multidimensionality [62]. The framework was developed by Hart and Staveland in 1988 [68], and it is used to assess workload on six 7-point Likert Scales. The framework was originally intended to be used for aviation, however, in 2006 Hart reflected on the use of the framework and found it to be more used in studies that focused on interface designs and evaluations [69]. The framework consists of six sub-scales that measure factors associated with completing a task:

1. Mental demands: how much mental and perceptual activity was required?
2. Physical demands: how much physical activity was required?
3. Temporal demands: how much time pressure occurred?
4. Performance: how successful do you think you were in accomplishing the goals of the task set by the experimenter?
5. Effort: how hard did you have to work - mentally and physically - to accomplish your level of performance?
6. Frustration level: how insecure, discouraged, irritated, stressed versus secure, content, and relaxed did you feel during the task?

Combining the results of the six sub-scales will achieve a measure of mental workload (cognitive load) [62].

In 1992, Paas came up with a new tool for measuring mental workload, the mental-effort rating scale [63]. Subjects had to report their invested effort in this 9-point Likert Scale, transforming the amount of mental effort into tangible numerical values. In Paas's study, a correlation between self-rated mental effort and test performance was found. Users who were presented with a task with easier-to-understand instructions experienced lower cognitive load, received better learning outcomes and rated their mental effort lower than users who were presented with more difficult task instructions. Moreover, the more difficult task instructions imposed a higher cognitive load [63]. The 9-point scale has later been deemed highly reliable [70].

Continued work on the 9-point Likert Scale by Paas, has been conducted by Ouwehand Et al. titled "*Measuring Cognitive Load: Are There More Valid Alternatives to Likert Rating Scales?* [71]. In this article, they investigated the validity of four subjective rating scales that differed in visual appearances. As a result of the rating scales, the mental effort was measured. The four scales were: the 9-point Likert Scale, the Visual Analogue Scale (0-100%), and two pictorial scales, one consisting of emoticons ranging from a blue-smiley (relaxed) to a red-smiley (stressed), while the last scale depicted as nine weights ranging from 1-9 kilogram. It was found that the numerical

scales suggested a better reflection of underlying complex problem solving, while the pictorial scales reflected the underlying simple problem solving [71]. This becomes an interesting subject to delve into as using a variety of measurement scales could help in measuring the mental effort in using the augmented reality application. At this stage of the project, we do not know how challenging or straining the application is. Because of this, we must take the findings of Ouwehand's study into account. However, the challenge for us, the developers, lies in distinguishing between what a complex and a simple problem is. We must make this distinction in order for the users (or participants) to provide us with the most optimal subjective measurements. One thing is that we present the users with a technology that in most cases can be quite abstract, and this together with not succeeding in helping them could result in sub-optimal results. Likewise, we must consider our wording carefully, as it was found that *mental effort* and *difficulty* are related, and failing to use appropriate wording can result in ensuing consequences in the results [62][72]. For example, users may encounter a problem so difficult that they are unable to make any realistic effort. Although correlated, the two terms do not always match [62].

The framework-of-choice is based on multiple conditions. 1. The framework outputs useful and detailed analysis of cognitive load. 2. The framework must be easily integrated into the testing framework. 3. The framework should consider task differentiation and individual cognitive load output. Having examined the two frameworks it was found that both frameworks do output useful and detailed analysis of cognitive load. NASA-TLX provides a more multidimensional workload assessment as it focuses on mental, physical, and temporal demands as well as effort, frustration, and performance. Paas's mental-effort rating scale (or 9-point Likert Scale) provides a more direct and simple measure of cognitive load based on mental effort. Subsequently, both framework ticks the box of the first condition. The second condition is much more subjective to the project and its testing framework. Also what is "easily integrated" can depend on who you ask. In terms of this project, we consider it to be "easily integrated" if: we do not have to make considerable adjustments to the prototype as well as not increasing the duration of the testing session substantially. Both frameworks are intended to be performed after a user (or participant) has ended a task. After the completion of a task they will be asked to fill in a short questionnaire revolving either their Task Load Index (NASA-TLX) or Mental Effort (Mental-Effort Rating Scale). Accordingly, there was not found a large difference in the two frameworks in terms of how easy they can be integrated. In addition, both methods are found to be equally time-consuming as you must, in both procedures, interfere with participant's after every ended task. That being stated, the NASA-TLX framework's questionnaire consists of six sub-scales, each containing one question the participants must answer. Oppositely, Paas's questionnaire only consist of one question. Exposing the participants for only one question rather than six will presumably reduce the duration of the experiment by a few minutes, helping to lessen the already cognitive straining assignment. Lastly, the framework-of-choice must consider task differentiation and individual cognitive load output. As NASA-TLX comes equipped with six sub-scales across different dimensions, it offers a com-

prehensive understanding of the overall workload experienced by individuals. Furthermore, it provides knowledge on how a participant has rated their cognitive workload for each task as well as the possibility of calculating an overall score. Likewise, Paas's Mental Effort Rating Scale also provides a single score for each task and averaging and calculating the mental effort mean score across the participants would result in a total mental effort score. The only difference between the two methods, is the more in-depth analysis due to NASA-TLX's multidimensionality.

Based on these findings our framework-of-choice when it comes to measuring cognitive load will be the NASA-TLX framework. As just proven it provides the multidimensionality that the Mental Effort Rating Scale lacks. Although, the NASA-TLX framework does seem to be more time-consuming, due to its six sub-scales, it does not change the fact that both frameworks interfere with the participants. Moreover, as the participants already are asked to fill in a questionnaire with one or more items, we see this as an opportunity to gather additional information about their experience. Furthermore, this information can also be used to not only output information about their cognitive load levels, but also provide an indication of their overall user experience.

2.3.4.2 Behavioral

Behavioral measures refer to the collection of data related to observable actions and responses in a particular context. In the book *Robust multimodal cognitive load measurement* [61], it is stated that response-based behavioral features are those that can be extracted from any user activity. They provide examples of eye-gaze tracking, mouse pointing and clicking, keyboard usage, and gait patterns. We will primarily focus on eye-gaze tracking, and learn to understand how it can be used for this project.

When developing for human-computer interactions it becomes vital to understand how users interact with interfaces to improve usability and design. For this topic, eye-gaze tracking can be used. Eye-gaze tracking is a technology that is used for monitoring and recording the eye movements of individuals, providing information about where an individual looks, how their gaze moves, and their fixate time on points of interest. Gütl Et al. [73] found that visual functions are partly involuntary as the eye can be drawn to salient items in the visual field. However, the act of gazing is under complete control and as a result, can be considered a behavioral measure [61]. To incorporate this measurement into the study the usage of devices like Apple Vision Pro, Microsoft HoloLens (see Section 2.1), or even a standard webcam could be used for obtaining this data. However, as we make a great effort in trying to diminish the usage of cumbersome headsets or devices and the difficulties of monitoring the eyes through a standard webcam due to the natural movement of augmented reality applications, it becomes difficult to make use of this aspect of behavioral measures. Additionally, as we intend to use physiological measures the complexity of the study rises if we include behavioral measures as well. The greater amount of measures, the more difficult the study and its variables become to interpret. Managing multiple

measures may lead to confusion or misinterpretation of results. Additionally, it can lead to Type I errors, especially if correlations for multiple comparisons are not considered appropriately. As a result, the null hypothesises could be rejected, even though it very well could be true. Although this should not be the sole reason for not using behavioral measures, it correlates with the already existing complexity of utilizing head mounts or other devices that you do not have access to in a real-world environment. That being stated, for future studies, examining behavioral aspects of the augmented reality application could be of interest, but for this project, it stays out of scope.

2.3.4.3 Physiological

Physiological measures can be used to evaluate user experience, especially within the immersive context of augmented reality (AR). Physiological measures, such as heart rate, electroencephalogram (EEG), or electrodermal activity (EDA) offer an objective method for evaluation, which can be used to assess the user's cognitive and emotional responses to the AR environment. Due to our previous experience with physiological measures, it would be considerably easier to integrate a physiological measure into our evaluation framework [14][74]. Physiological signals like heart rate, skin conductance, or facial expression can be monitored through relatively non-intrusive wearable sensors to assess the user's emotional state or stress levels. This would allow the researchers to know which specific tasks and interactions resulted in more cognitive load or stress for the user. EEG sensors can be used to indicate a user's cognitive load, but are more intrusive and require an advanced testing setup. In the report *Measuring cognitive load in augmented reality with physiological methods* by Yuko Suzuki it is mentioned that physiological measures require a baseline reference to be an effective measure. They are usually used to complement more conventional methods such as self-report measures [75]. The study found that the most common self-report measure used with physiological data is the NASA-TLX framework. NASA-TLX is perceived as a reliable tool for measuring workload, which makes it a strong reference for physiological data. In Suzuki's report, it is furthermore highlighted that raw physiological data will only provide meaningful insights if they are supported by contextual benchmarks. For example, a heightened heart rate could indicate increased cognitive load or stress, but without benchmarks for normal versus elevated heart rate levels, the interpretation will be unclear. Furthermore, they suggest the importance of distinguishing between cognitive loads (intrinsic, extraneous, germane) that are a result of technological AR aspects or whether they stem from task complexity [75]. Another way to increase the quality of the physiological data is to differentiate between task complexity and the expertise level of the test participant, by comparing cognitive load levels during testing. For instance, a task that seems complex to a novice AR user might be easy and manageable for a user who is an expert in AR, leading to a difference in cognitive load when comparing the participants [75].

The assessment of cognitive load can be significantly improved by using physiological measures to support self-report measures [76]. Studies have acknowledged the limitations of purely us-

ing self-report measures. People tend to under or over-report their mental effort depending on various tasks [56][76]. Physiological measures such as heart rate or EDA could offer a more objective way to assess cognitive load. The report *Let Complexity Bring Clarity: A Multidimensional Assessment of Cognitive Load Using Physiological Measures* by Emma J. Nilsson Et al., proposes a multidimensional assessment approach that combines physiological and self-report measures. By combining these methods, the researchers can hopefully capture a more complete picture of cognitive load. It is important to be aware that physiological measures can correlate with diverse psychological responses. This one-to-many association significantly narrows the scope of what can be learned from individual measures alone [77]. Consequently, although some physiological measures may correlate with cognitive load, they can not exclusively be regarded as a measure of cognitive load. Previous research indicates that the correlation between physiological measures and cognitive load are independent of each other, which means that one variable does not necessarily cause direct changes in the other variables, they simply share a relationship that can be observed in the data. This indicates that different physiological measures each capture unique aspects of how a person reacts to increased cognitive demands. It is therefore suggested that using multiple physiological measures in conjunction will result in a richer and more comprehensible view of cognitive load [76]. Some of the physiological measures that are typically used to assess cognitive load and have been known to provide significant results are heart rate (HR), heart rate variability (HRV), blink rate (EBR), skin conductance (EDA), and pupil diameter (PD). Another physiological measure that is commonly used, is EEG. However, EEG measures vary greatly in terms of which EEG channels are used as well as the data post-processing procedure. This makes the process of gathering data using EEG more complex than other physiological measures and since they do not provide more significant or correlated responses, it could make sense to steer around using EEG as a physiological measure [76]. In general, establishing a baseline is a crucial step when dealing with physiological data. For example, when assessing heart rate it is essential to measure the test participant's heart rate in a controlled and neutral environment before any AR interaction, in order to establish a baseline for future reference. This reference point will be analyzed and compared with all subsequent heart rate measurements. The need for baselines is caused by the natural variability of physiological responses among different individuals, which can vary greatly depending on the participant's age, fitness level, health, emotional state, stress levels, etc. [78].

Under stressful conditions, the human body can display multiple physiological changes such as elevated heart rates, faster breathing, and increased sweating. The devices presented below are capable of measuring these physiological features. However, each device comes with its own pros and cons. One of the more common methods for measuring heart rate (HR) or heart rate variability (HRV) includes photoplethysmography (PPG) sensors which detect blood volume pulse changes (BVP) in the skin, which can indirectly reflect heart rate or heart rate variability. It utilizes a light source and a photodetector at the surface of the skin to measure the absorbance of light, which

varies depending on a person's blood pulse. The technology is widely used in consumer products like smartwatches and fitness bands as it allows for continuous and non-invasive monitoring of heart rate. While they might be convenient for personal use, their accuracy can be affected by motion, changes in light, and other factors that must considered, if using a PPG device in a scientific setting [79]. Another technique for measuring the heart rate of a subject is electrocardiography (ECG or EKG) devices. An ECG device directly measures the electrical activity of the heart using electrodes placed on the skin, typically around the chest area. They are more precise than PPG devices and are commonly seen in both clinical and research settings. ECG devices are generally more obtrusive compared to other heart rate monitoring methods such as PPG. This increased obtrusiveness is primarily due to the need for placing multiple electrodes across the chest and other body parts, which can be somewhat intrusive since it requires participants to lift up their shirts. This might not be a significant issue in a clinical setting, but in a work or study environment, this may make it significantly harder to recruit test participants. Also, considering the spatial nature of our AR application, where users are encouraged to move freely around, it is important to leverage non-intrusive physiological measures [80].

Another non-intrusive measure is electrodermal activity (EDA), also known as galvanic skin response (GSR), which measures the electrical conductance of the skin, which varies depending on the skin's moisture level. It is commonly used to measure psychological features such as emotions, affect, or cognitive load [81]. The sensors for measuring EDA are typically placed on the fingers or in the palm of the hand, which are areas with a high density of sweat glands. Measuring the participant's EDA under rest and establishing a baseline is a crucial step for the data to be useful later on as with heart rate. In a study by Buchwald et al. [81], it was explored whether or not EDA could specifically measure cognitive load and its three sub-constructs: intrinsic, extraneous, and germane. Their findings suggest that, while EDA can generally indicate cognitive load, it is unclear whether it can be used to differentiate between the sub-constructs of cognitive load.

By integrating physiological measures such as EDA and BVP into our methodologies, it will allow us to better capture user experiences and measure cognitive load. These non-intrusive devices facilitate free movement, which is crucial in a spatial AR application. Including EDA and BVP, will enable us to obtain a more nuanced understanding of how physiological responses correlate with user interactions in a real-time AR environment. Moving into the examination of task performance, these physiological measures will be crucial when analyzing the task performance data. By comparing the physiological data with task events and performance, it will be easier to understand the implications of the AR design on user experience. This approach will allow us to identify specific AR features that cause an increase in cognitive load, which can directly impact task performance.

2.3.4.4 Task Performance

Fang Chen Et al. [61] states that *"learning is hindered when working memory capacity is overloaded, and therefore a drop in performance will be the result of an increase in overall cognitive load."* [61][66], and in order to measure this, you can use task performance. Task performance is a quantitative measure used to assess how successfully users can complete pre-determined tasks. This typically involves features such as completion time, error rate, click/taps, and whether the end result of the tasks was correct [82]. Along with these measures, Fang Chen Et al. [61] also presents the dual-task paradigm technique which concurrently evaluates an individual's performance in a secondary task while assessing the cognitive load allocated to the primary task. Although this technique is widely used in the field of task performance and cognitive load, it may not fit into the scope of this project. The current focus is not on examining what tasks require more cognitive resources. Instead, we want to examine how the cognitive resources react when a user interacts with the AR application. Due to this, we will consider to present a user with one task at a time, resulting in the dual-task paradigm being superfluous. Other task performance measures we scope in on are: *Task on Time* and *Task Success Rate*. We have previously touched upon these measures of task performance (see Section 2.3.2.5), and found them to be commonly used for measuring user experience. However, these measures can also be converted to cognitive load measures. Time on Task can help to determine the difficulty of a task. Especially, if you assume that more time spent on completing a task equals a more mentally demanding task. This would indicate that higher mental levels are required and as a result high cognitive load levels were used to complete the task. In contrast, if a user quickly completes a task, it indicates that the task was easy to understand and did not require much mental capacity. Likewise, Task Success Rate can support the evidence of the Time on Task measure by looking for a decrease in task success rate that being the number of incorrect task answers. This negative task success rate would imply higher levels of cognitive load required for completing the task successfully. If a user's mental capacity becomes overwhelmed, they might find the task more challenging and more errors could occur as a result thereof [61].

2.4 Final Problem Statement

This far, we have sought to understand the interplay between augmented reality (see Section 2.1) and data visualization (see Section 2.2). We have used this knowledge to further build upon the idea of creating an AR application for 3D data visualization. It has become clear that utilizing an iOS device is suitable for augmenting data visualization on to a real-world environment, however, we cannot yet prove that augmented reality optimizes and enhances the user experience in 3D data visualization. As a result, we have not found the exact solution for our initial problem statement (see Chapter 1), but examining the field of work has provided us with additional knowledge and new questions to answer. Reviewing the work of Norman (see Section 2.3.2) and Sweller (see Section 2.3.3) provided us with knowledge on how to develop a product specifically tailored

to users and further yielded knowledge on the cognitive load and resources. This is where the project becomes twofold. One aspect is to develop and enhance a user's user experience in data visualization, and the other aspect is to measure a user's cognitive load while utilizing the application. Based on the findings throughout the Analysis Chapter (see Chapter 2), we opted to split the final problem into two research questions. The research questions are as follows:

Research question 1:

How does an augmented reality application for data visualization enhance the user experience compared to a static version?

Alternative hypothesis - H_1: Data visualization in augmented reality enhances the user experience when compared to a static version.

Null hypothesis - H_0: Data visualization in augmented reality does not enhance the user experience when compared to a static version.

Research question 2:

By use of a multimodal cognitive load measuring approach, how does an augmented reality application for data visualization influence extraneous cognitive load?

Alternative hypothesis - H_1: An augmented reality application for data visualization does influence extraneous cognitive load.

Null hypothesis - H_0: An augmented reality application for data visualization does not influence extraneous cognitive load.

3 Methodology

In detail, the research approach used in this study will be presented. This includes insight into the testing plan, participants, and methodologies. The chosen procedures are based on the findings in the analysis chapter. See Chapter 2 for more details.

3.1 Participants

All participants for this study were from the Technical Faculty of Aalborg University in Copenhagen. In each of the first, second, and third iterations a total of eight participants were accumulated. As part of the foundation (see Table 1) the participants were asked to state their knowledge in statistics. This knowledge was used to determine their expertise allowing insights into each participant's statistical perspective on the application. The interviews were later transcribed verbatim and then analyzed. Building upon the foundation it was later found that the statistical level of the users was not as important for the study as their 3D visualization skills. Consequently, the project pivoted to focus on the user's familiarity with 3D visualization rather than their statistical knowledge. In the final iteration, a total of 58 participated. All participants gave informed consent, and they were informed that they could withdraw their data from the study at any time. Furthermore, all participants were provided with anonymized ID numbers to protect private or sensitive information.

3.2 Iterative Design and Prototyping

The prototype development was within an iterative process. The prototype was designed, developed, and evaluated a total of three times before the final testing session. This ensured a functional product. This was based on Don Norman's methodology, User-Centered Design (see Section 2.3.2). The prototyping included interviews and usability testing. Three researchers and two students at Aalborg University in Copenhagen were interviewed as part of the preliminary work. The interviews were conducted to identify their preferences when working with and visualizing statistics. This was later used as the user requirements and success criteria for the product. The various usability tests were assessed by having participants fill in the System Usability Scale questionnaire. Pilot testing was conducted internally prior to the final testing session. The final testing protocol was tested, to ensure consistency and proficiency in equipping the physiological equipment and obtaining data. The protocol included information only available to the researchers. See 10.3 Appendix C for the full protocol. Between groups were used for the final testing session. The experimental group used the application dynamically, meaning they were able to move freely around while examining the datasets. The control group was static while using the application, meaning they were limited to using tap gestures for manipulation (scale, rotate, translate).

3.3 Evaluation and Data Analysis

In this subsection, we will explain the methodologies employed to process and analyze data obtained from multiple sources, which will include self-report measures, physiological data, and task performance measures. This section will discuss the instruments used for collecting physiological data, the various preprocessing steps, and the tools used throughout these processes. The final evaluation of the augmented reality application was conducted using two questionnaires: NASA-TLX and UEQ. Additionally, physiological measures will also be collected, specifically Electrodermal Activity (EDA) and Blood Volume Pulse (BVP). Lastly, we will also gather task performance measures such as Time on Task and Task Success Rate. Participants will be required to fill in the NASA-TLX questionnaire after the completion of each task, resulting in a total of five completions per participant. Additionally, the User Experience Questionnaire (UEQ) will be administered at the end, after the completion of all tasks. This multifaceted evaluation will aim to correlate the various measures with each other to provide insights into user experience and cognitive load (more specifically, extraneous cognitive load.).

3.3.1 Self-Report Measures

For the evaluation we will utilize two self-report measures: the UEQ and NASA-TLX questionnaires to assess user experience and cognitive load. The NASA-TLX (Task Load Index) will be used to assess cognitive load and the participants will be asked to fill in the NASA-TLX questionnaire after completion of each task, resulting in a total of five Task Load Index's per participant. The UEQ data will be collected after completion of all tasks, and used to assess user experience.

3.3.1.1 User Experience Questionnaire (UEQ)

For measuring user experience, we will use the User Experience Questionnaire (UEQ). As mentioned in Section 2.3.2.5, it consists of 26-items in a 7-point Likert Scale questionnaire. To have more quantifiable numbers to work with, the 7-point Likert scale responses will be transformed into values ranging from -3 (most negative) to +3 (most positive). This is suggested by Dr. Martin Schrepp, one of the UEQ developers. The six sub-scales have got different items attached to them with *Attractiveness* (6-items) having 6-items, resulting in it being the larger sub-scale, while the remaining five sub-scales have a total of 4-items attached each: *Perspicuity* (4-items), *Efficiency* (4-items), *Dependability* (4-items), *Stimulation* (4-items), and *Novelty* (4-items). Each of the six sub-scales for each participant will then be averaged, which will provide a numerical value between -3 and +3. This data transformation must be done for both the experimental and control group, leaving us with sub-scale means of each of the six sub-scales and additional measures like: standard deviation, confidence levels, and confidence intervals. In order to determine whether there are significant differences between the experimental and control groups UEQ responses, we will perform a two-sample t-test on each groups mean sub-scales. The two-sample t-test will help in identifying whether the observed differences in the user experience scores between the two

groups are statistically significant. We have set a significance level (alpha-level) of 0.05, meaning that a p-value below 0.05 will indicate a statistically significant difference, which is a commonly accepted threshold in many areas of science [83][84].

3.3.1.2 NASA-TLX (Task Load Index)

For measuring cognitive load, we will use the NASA-TLX questionnaire. As mentioned in Section 2.3.4.1, the NASA-TLX is a widely used tool for assessing cognitive workload, and it does this across six dimensions using a 7-point Likert scale. These dimensions include *mental demands*, *physical demands*, *temporal demands*, *performance*, *effort*, and *frustration*.

As the participants will be asked to fill in the 7-point Likert scale, and the NASA-TLX originally being a scale from 0-100, we must convert the scale. This is done by transforming the users answers from 1-7 to a value between 0-100. We will do this by multiplying a users value with 100 and divide it by 7, converting the answer to a numerical value that fits the 0-100 scale. E.g. if a participant answers 3, it would be converted to a score of 42.85 ($3 \times (100 \div 7)$). After converting the values to fit the 0-100 scale, each of the six sub-scales must be assigned a weight. For this project the sub-scale weights are as follows: Mental Demand: 0.3, Physical Demand: 0.1, Temporal Demand: 0.1, Performance: 0.1, Effort 0.3, Frustration: 0.1. You calculate the weighted scores by multiplying each sub-scale rating by its corresponding weight. E.g. 42.85×0.3, leaving you with a weight rating of 12.85. Transforming the data from the 7-point Likert scale to the 0-100 scale and calculating the weighted scores must be done for each individual in both groups. Calculating the sum of a sub-scale's weights and dividing that sum with the total number of participants (n = 29), leaves you with the average weight score for that category. Doing this for each of the six sub-scales and adding them together provides the overall NASA-TLX score for one of the five tasks. Consequently, this procedure must be done a total of 10 times (one time per task per group), resulting in one overall NASA-TLX score per task. This will provide a Task Load Index for each of the five tasks in both groups, in which we can compare and examine which group reported the largest mental workload.

3.3.2 Task Performance Measures (Time on Task and Task Success Rate)

Task performance will be evaluated through two primary measures: Task on Time and Task Success Rate. Time on Task measures the duration taken by participants to complete each task, while Task Success Rate assesses the effectiveness of the participants in accomplishing the tasks.

3.3.2.1 Time on Task

Time on Task will be calculated by leveraging the time series data from the physiological measuring devices, which timestamps the beginning and end of each task. This method will allow us to precisely determine the duration each participant spent on each task. It is important to note that task time alone does not directly indicate cognitive load, as individuals have varying speeds of

task completion due to factors such as familiarity with the task, stress levels, and personal conditions. Additionally, external factors such as the testing environment, participants health, and mood can also influence task time and overall performance. Although, more complex tasks will typically take longer to complete, the relationship between time and complexity is not always linear. Task time will be used to normalize some of the physiological measures later in the Discussion (see Chapter 8), providing a more nuanced understanding of the participants' cognitive load and performance.

3.3.2.2 Task Success Rate

Task Success Rate assesses the effectiveness of the participants in accomplishing the tasks. Each task had specific criteria for successful completion, and participants' performances were evaluated against these criteria. The success rate provided a clear measure of proficiency, indicating how well participants could perform the tasks using the augmented reality application. High success rates would suggest that the application is intuitive and easy to use, while lower success rates might indicate areas where the application could be improved. Additionally, a higher value of success rate could suggest, especially within an individual user, that lower extraneous cognitive load levels were used during the task. In contrast, if a user answers wrongly on multiple tasks it could indicate that higher levels of extraneous cognitive load is used, as the tasks are more difficult to that individual user. That being stated, as the participants will be randomly assigned to the two groups (experimental and control), the users proficiency will be randomly allocated between the groups. However, looking closer on a individual's Task Success Rate could show different levels in intrinsic knowledge. Moreover, we had a great influence in determining the intrinsic difficulty levels of each task, meaning that we could exclusively establish a balanced array of tasks ranging from easy to hard. This would ensure that the success rate would not be too low nor high, if the tasks were overtly easy or hard respectively.

Utilizing Task Success Rate and combining it with Task on Time and physiological data, will offer a comprehensive view of user performance and cognitive load, allowing us to identify patterns and areas for improvement in the application design.

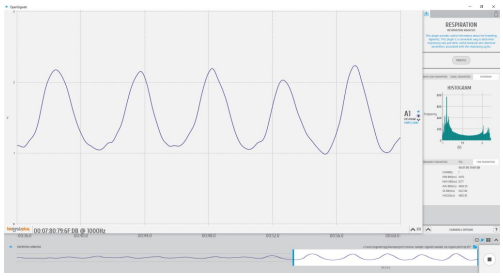
3.3.3 Physiological Data

The final evaluation will utilize two physiological measures: Blood Volume Pulse (BVP) and Electrodermal Activity (EDA). As previously stated in the analysis (see Section 2.3.4.3), these measures will be collected to provide objective data on participant reactions during task performance. In this subsection, the tools used to collect, process, and analyze BVP and EDA data will be described.

The physiological data for this study will be collected using devices from PluxBioSignal, a company known for its advanced biosignal acquisition technology [85]. The PluxBioSignal system includes sensors, an amplifier/hub (see Figure 9a), and accompanying software designed to pro-

vide accurate and reliable measurements of physiological responses (see Figure 9b). The software will handle the raw sensor input and convert it into meaningful values, and can be outputted as a file or streamed real-time over a TCP/IP connection or using LabStreamingLayer (LSL) protocol [85][86]. The PluxBioSignal system used in this study consists of a 4-channel hub that collects and digitize the signals from the connected sensors and transmits them via Bluetooth to a computer for real-time recording and visualization using the OpenSignals (r)evolution software. The software's real-time visualization feature will be very useful for the final test, as it will allow us to track the signal and potential issues while the test is on-going. The BVP and EDA sensors from PluxBioSignal kit are connected to the hub, which automatically detects and configures them. The PluxBioSignal amplifier/hub supports up to a 3000Hz sampling frequency, but in this study, the default 1000Hz sampling frequency will be used. It supports 16-bit resolution per channel and has a battery life of approximately 10 hours [85]. The software offers various add-ons for specific analyses, such as heart rate variability and electrodermal activity event analysis, however, we will perform these analyses ourselves using Python and open-source libraries, as this allows us to control the processing of the data, and is essentially cost-free.

(a) The 4-channel amplifier/hub used to collect and digitize BVP and EDA signal [87].



(b) The OpenSignals (r)evolution software displaying real-time physiological data collected from PluxBioSignal sensors during a session [88].

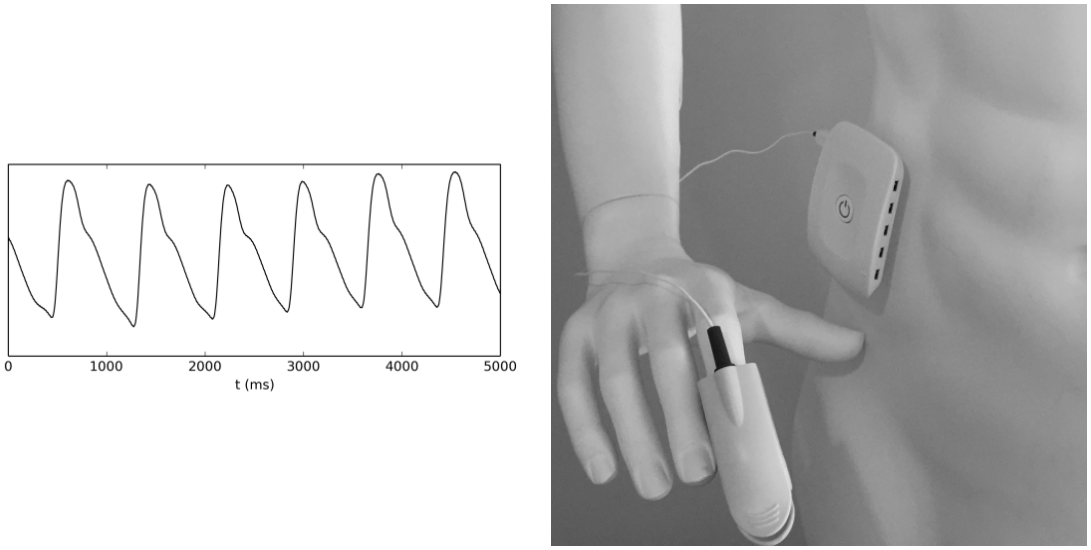
Figure 9: Shows the 4-channel amplifier/hub (a) and OpenSignals (r)evolution software (b).

In the final evaluation, the physiological measures for each task will be tested for a normal distribution. If the measure for a task is normally distributed, we will use a paired t-test to determine whether the difference between the experimental and control groups is truly significant or due to randomness. If the data is not normally distributed, we will use the Mann-Whitney U test. To be considered truly significant, the data must have a p-value below 0.05.

3.3.3.1 Blood Volume Pulse (BVP)

As mentioned in the Analysis (see Section 2.3.4.3), BVP measures changes in blood volume within the vascular bed of tissue and can provide insights into heart rate and heart rate variability (see Figure 10a). The BVP sensor used in this study is a non-invasive optical sensor by PluxBioSignals that detects changes in arterial translucency with each heartbeat. This sensor has a plastic clip-on housing for easy placement on the finger, which should minimize interference from external light

sources [89] (see Figure 10b). The BVP device will be placed on the index finger of the test participant, specifically on the participants non-dominant hand, as they will be using their dominant hand to carry and interact with the iPad. The device is plugged into one of the four available ports on the PluxBioSignal hub. As previously mentioned, BVP devices are very sensitive to light, so it is important that the light settings in the testing environment are controlled and that lighting in the room is as similar as possible across different testing sessions, which will be conducted over multiple days. Furthermore, the BVP sensor can be sensitive to movement, particularly erratic movements, so participants will have to treat the device carefully.



(a) Typical raw BVP signal collected using PluxBioSignal sensors, illustrating the waveform of blood volume changes over time [89].

(b) Placement of the BVP sensor on the index finger of a participants hand, connected to the PluxBioSignal hub for data collection [89].

Figure 10: Shows a raw BVP signal (a) and the BVP sensor connected to an index finger (b).

Specifically, the BVP data will be processed to derive metrics such as Beats Per Minute (BPM), which indicates heart rate, and Standard Deviation of Normal-to-Normal intervals (SDNN) and Root Mean Square of Successive Differences (RMSSD), which are measures of heart rate variability. The following measures will be derived using the HeartPy library, a Python toolkit for heart rate analysis. The *heartpy.process()* function will be used to pre-process and apply filters to the raw BVP data, removing noise and artifacts in the data and converting it into BPM and a wide range of heart rate variability measures [90].

These are the following measures the raw BVP data will be converted into:

- **Beats Per Minute (BPM):** Beats Per Minute (BPM) is a measure of heart rate, which represents the number of heartbeats per minute. To interpret BPM results, a baseline will be established for each participant, which will then be subtracted from the individual task-related BVP data and converted into a percentage to normalize it and highlight the differ-

ences caused by the tasks. This normalization is highly important as heart rate measures, specifically BVP, can vary greatly depending on the physical and mental condition of the participant. Higher BPM values, shown as the percentage increase relative to the baseline, can indicate higher physical or mental stress levels, whereas lower BPM values typically reflect a more relaxed state [91].

$$\text{BPM Percentage Increase} = \left(\frac{\text{Task BPM} - \text{Baseline BPM}}{\text{Baseline BPM}} \right) \times 100$$

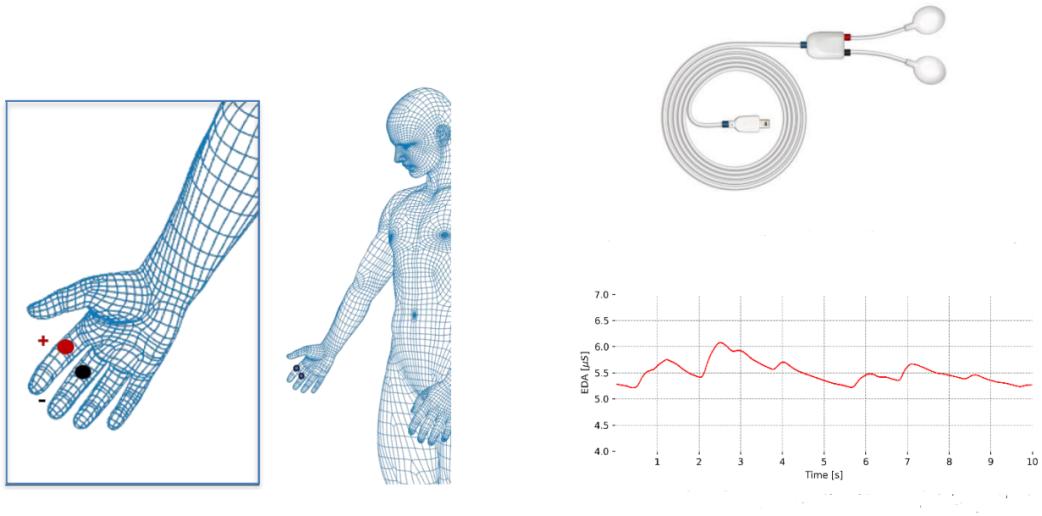
- **Standard Deviation of Normal-to-Normal intervals (SDNN):** SDNN measures the variability in time between successive heartbeats (normal-to-normal intervals). SDNN is a common measure for assessing heart rate variability, where higher values typically indicate generally better cardiovascular fitness and autonomic flexibility. Lower SDNN values can usually indicate reduced variability, which is associated with stress, fatigue, or poor autonomic function [92]. The conventional time needed to properly record SDNN is a minimum of five minutes, which is likely a lot more than the time needed to complete the individual tasks in our final analysis. Since we will be calculating the SDNN for each task interval, it is noteworthy that some research has found that short-term recording periods from 60 to 240 seconds are also viable, which likely fits a lot more with the time range of our final test tasks. SDNN is actually considered the standard benchmark for assessing cardiac risk in medical settings, particularly when recorded over a 24-hour period and can predict mortality. In a longer recording, SDNN values below 50 indicate poor health and values above 100 are classified as healthy [93][94], however these are drastic generalizations.
- **Root Mean Square of Successive Differences (RMSSD):** RMSSD is a measure of heart rate variability that focuses on short term changes in heart rate. It mainly reflects the activity of the parasympathetic nervous system, which is responsible for the 'rest and digest' functions of the body. Higher RMSSD values indicate better parasympathetic activity and good autonomic control, while lower values can suggest reduced parasympathetic activity and potential stress or imbalance. RMSSD values below 20 may indicate poor autonomic regulation, while values between 20-89 are generally considered healthy. While a minimum of 5 minutes of recording is typically used, some researchers have suggested that very short term periods of 10 to 30 seconds also could be viable [93][94]. Although these shorter periods may not be compatible with the time some participants will spend completing a task, they can still provide different and meaningful insights into heart rate variability across the various tasks.

3.3.3.2 Electrodermal Activity (EDA)

As mentioned in the Analysis (see Section 2.3.4.3), EDA tracks changes in skin conductance, which is an indicator of sweat gland activity which has been closely linked to emotional arousal (see Fig-

ure 11b). An EDA sensor measures the electrical properties of the skin that change with sweat gland activity, providing a direct measure of the sympathetic nervous system activity, which is responsible for the 'fight or flight' response [95]. EDA responses can be influenced by numerous factors such as ambient temperature, humidity, participants hydration level, emotional state, and recent physical activity [96][97][98]. EDA responses are typically categorized into two components: tonic and phasic. The tonic component (skin conductance level, SCL) represents the slow changing baseline of skin conductance which happens over time. The phasic component (skin conductance response, SCR) is instead temporary changes triggered by specific stimuli or events [98]. In this study, we have decided to focus on SCR specifically, due to its immediate indication of sympathetic nervous system activity [99]. While SCL would provide meaningful information on the overall arousal state of the test participant, it usually requires longer recording times as the value usually takes 10-30 seconds to change [100], which is close to how fast we estimate some participants might complete a task.

The EDA sensor used in this study is a non-invasive and lightweight sensor by PluxBioSignal with a high signal-to-noise ratio and medical-grade raw data output [101]. The PluxBioSignals datasheet explains that the first sensor (red) should be placed on the index finger, while the second sensor (black) needs to be positioned on the middle finger (see Figure 11a).



(a) The PluxBioSignals data sheet indicates that the red EDA sensor should be placed on the index finger, while the black sensor needs to be positioned on the middle finger [101].

(b) The top picture shows a drawing of the specific EDA sensor used in this project. The bottom image represents a raw EDA signal collected using the PluxBioSignal EDA sensor, illustrating skin conductance changes over time [101].

Figure 11: Shows the placements of the PluxBioSignals EDA sensor, the equipment, and an illustration of its raw signal.

The raw EDA data will be processed to derive SCR metrics such as SCR peaks count, SCR peaks

amplitude mean, and raw EDA mean. The following measures will be derived using the NeuroKit2 library, a Python toolkit for physiological signal analysis. The *nk.eda_process()* convenience function that will automatically preprocesses the raw EDA data, removing noise and artifacts. This function employs algorithms to detect and quantify skin conductance responses by using peak detection and signal decomposition techniques. The *nk.eda_analyze()* function then analyzes the processed signals, providing detailed metrics that reflect SCR activity by using peak detection and signal decomposition techniques [102].

These are the following measures the raw EDA data will be converted into:

- **SCR Peaks Count**: The number of SCR peaks that occur during a recording. Since participants will naturally require different amounts of time to complete each task, those who spend more time (120-240 seconds) on a task will experience more SCR peaks than participants who complete the task in a shorter duration (30-90 seconds). Therefore, the number of peaks will be divided by the time taken to complete each task to normalize the data [103].
- **SCR Peaks Amplitude Mean**: The mean amplitude of all the SCR peaks that occur during a recording, reflecting the intensity of the skin conductance response. Furthermore, it is suggested that the SCR amplitude mean can serve as an effective measure of sympathetic nervous system activity [104]. The values are represented as the percentage increase relative to the baseline, using the following formula:

$$\text{SCR Amplitude Percentage Increase} = \left(\frac{\text{Task SCR Amplitude} - \text{Baseline SCR Amplitude}}{\text{Baseline SCR Amplitude}} \right) \times 100$$

- **Raw EDA Mean**: The mean value of the raw EDA signal over a task/time interval, which provides a measure of skin conductance during the task as a whole, and can be compared to the baseline. The values are represented as the percentage increase relative to the baseline, using the formula below:

$$\text{EDA Mean Percentage Increase} = \left(\frac{\text{Task EDA Mean} - \text{Baseline EDA Mean}}{\text{Baseline EDA Mean}} \right) \times 100$$

The combination of self-report measures, physiological data, and task performance metrics will offer a comprehensive view of the test participants interactions with the augmented reality application. This detailed methodology ensures that the study can accurately assess both cognitive load and user experience.

4 User Research

Preliminary work was conducted to investigate and assess the fundamental conditions we were to follow for this study's prototype. To gain this information we conducted five individual semi-structured interviews. The participants consisted of three researchers and two students from the Technical Faculty of Aalborg University in Copenhagen.

4.1 Interviews

Before the interviews could begin, the interviewees' had to give informed consent. They consented to the allowance of recording the interview and for it to later be transcribed, and used in the study.

4.1.1 Interview Procedure

The interview began by having the interviewees rate their statistical knowledge. They could choose between three levels (see Table 1). They would be categorized to a level depending on what statistical methods they tend to use in their work. For example, if they primarily use descriptive statistics they would be categorized to be in the fundamental level range, however, if they primarily use multivariate analysis they would be categorized to be in the intermediate level, and so forth. The interview guide used for the interviews can be found in 10.1 Appendix A, and the transcriptions of the five interviews can be found in 10.2 Appendix B.

Fundamental Level (Descriptive statistics, Basic probability concepts, Hypothesis testing)
Intermediate Level (Regression analysis, Experimental design, Multivariate analysis)
Advanced Level (Time series analysis, Bayesian statistics, Machine learning algorithms)

Table 1: The three statistical knowledge levels used for labeling the participants.

4.1.2 Interview Analysis

The interviews revolved around the interviewees statistical usage. Although conducting five interviews, despite not being a sufficient amount of interviewees to identify preferences (see Section 2.3.2.1), much can still be derived from the interviewees' answers, as the five interviewees were evenly split across the three knowledge level categories (fundamental (2), intermediate (1), and advanced (2)). It was found that the five interviewees, although varying in statistical levels, primarily would use Python and Excel for processing their data. In addition, we wanted to know what visualizations they found themselves using most often. Histograms, scatter plots, and box plots were the plots that were mentioned the most. Based on this, it can be argued that we should aim our focus towards these visualizations as they, based on the interviews, are the most used ones. However, it must be considered if it makes any sense to draw the plots in augmented reality. If the visualizations do not provide any advantages over the traditional 2D and 3D visualizations,

then the whole purpose of augmented reality becomes cumbersome and to some degree redundant. One interviewee mentioned this when asked about how augmented reality could enhance their data visualization workflow. They stated that using augmented reality for data with three or more dimensions "*(...) will be very useful, but if it's for normal correlation data or box plots or anything like that, I don't think it will add anything.*" (see 10.2.5 Appendix B). This is important to consider and must be considered in later iterations. Nevertheless, the idea of an interactive data visualization tool was found to be helpful in data visualization, especially with the possibility of real-time manipulation. It was thought to provide "*(...) more perspective*" to examine the heights of a dataset in a real-life environment (see 10.2.1 Appendix B). Additionally, "*(...) it could be more immersive* (see 10.2.3 Appendix B), especially if being able to "*(...) manipulate data - zoom in, zoom out 3D rotations.*" (see 10.2.5 Appendix B). Furthermore, the interviewees were asked to put some words on how they would measure the success of an AR data visualization tool, and prioritize their criteria from most important to least. It was derived from the interviews that ease of use was the most important aspect to consider. Second, was how easy to understand the software and data visualization for own use was, as well as presenting it to others. This goes well in-hand with the five interviewees finding data visualization in AR to work great for presentations of datasets, but it greatly depends on the context of use. If only presented through a device, it would not work with multiple people having to observe the data presentation. However, screen-sharing from the device to a larger display could solve that issue, which further would result in more people being able to observe the data simultaneously with interactions happening. This aspect of data presentation or interaction between multiple people at the same time, would not likely be a high-priority criteria in context of this project as this largely would increase the scope of the project. Lastly, the five interviewees mentioned learnability. Creating a new application with a new interface that must be learned to produce work, should have a short and not too steep learning curve. This must be considered and examined through iterative work and testing in the later iterations.

4.1.3 Success Criteria

As mentioned in Section 2.3.2.2, analyzing the interviews would help to form a set of requirements that will act as the success criteria of the product. The success criteria will be evaluated in later Chapters (see Chapters 5 and 8). They are as follows:

1. The application should be easy to use.
2. The data visualization should be easy to interpret.
3. Learning the interface and functionality of the application should be easy and simple to do.

5 Design and Implementation

As part of designing and developing the prototype for use in the final test, we opted to develop the application in an iterative process as previously mentioned in Chapter 3.2. The following iterations were each developed based on preliminary design choices grounded in User-Centered Design (Section 2.3.2), as well as feedback, observations, and data gathered throughout the iterations. This ensures that the application does not end up with any glaring issues and instead strives to be as intuitive and optimal for the user as possible. It is an important factor to limit the amount of errors, issues, or sub-optimal features as these can cause unintentional strain for the user and in turn influence the integrity of the final experiment with confounding variables.

5.1 The First Iteration

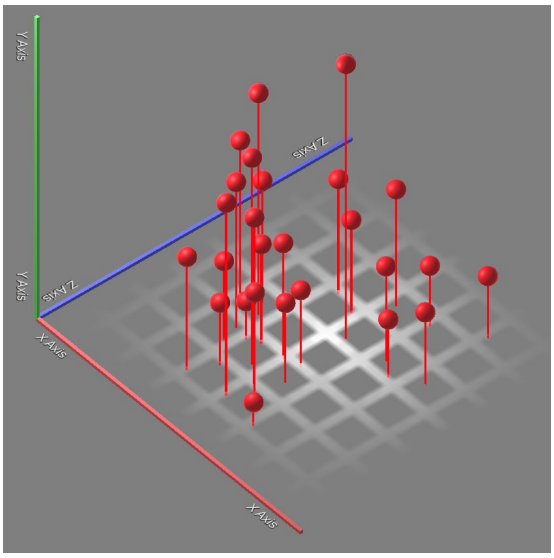
In this section, we will cover the steps taken from the design and implementation to the usability evaluation of the first iteration of the data visualization application. This iteration will be based primarily on many of the fundamental features and be rather exploratory as opposed to future iterations where the scope will be more narrow and specific. The main intention of this iteration is to test usability and to explore the users' experience.

5.1.1 Development

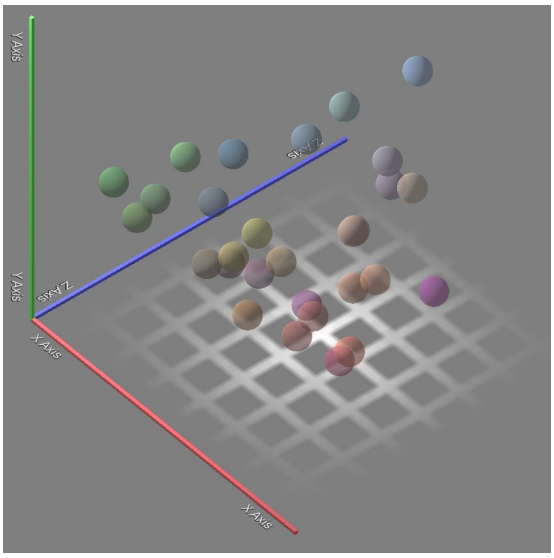
Many aspects were to be considered when starting the design process of a broad development scope, such as this. While there are many different directions and probable solutions for many of the different difficulties that we face, the intention is for us to best as possible reach a prototype for an initial first iteration usability test. This initial first iteration prototype should include some select essential features we would like to evaluate, as well as all the implementation of the fundamental features such as device integration and plane detection. A few of these select features for this prototype would include the placement of a dynamically scaled plot, a simple navigation menu feature to switch datasets in the plot, and lastly ensuring plot implementation is compatible with future dataset implementation using a unit cube. Finally, all design decisions should be considered around, and adhere to the design principles in UCD (see Section 2.3.2.3). The design and implementation of this iteration will be divided into three parts:

1. The technical device specification and AR integration (plane detection).
2. The dynamic plot placement with scalable length, width, and height.
3. The actual plotting of datasets into the dynamic 3D plot.

As part of the data plotting, we also want to have multiple different dataset visualizations, which the participants as part of the test should switch between. This would give insight into the impact of the visibility aspect of the visualizations. The dataset plots used for this iteration are illustrated in Figures 12 and 13.

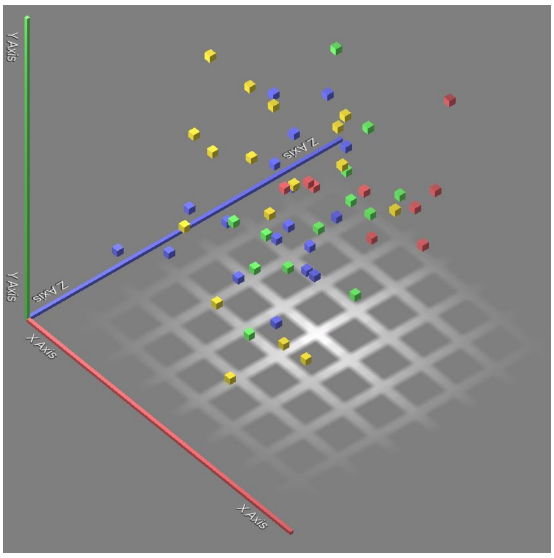


(a) Plot 1 - Medium sized red spheres with red vertical lines going to the floor.

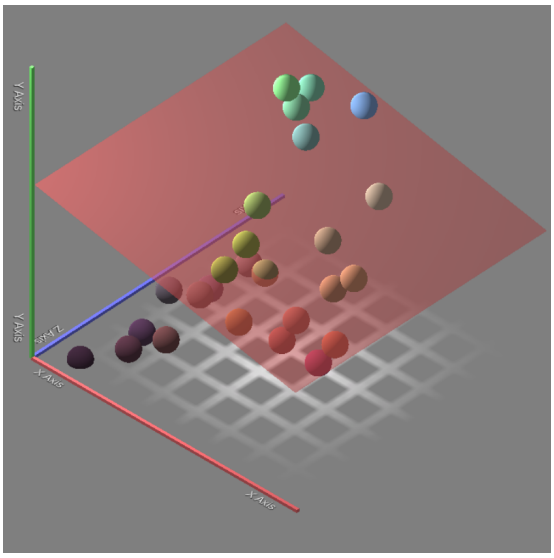


(b) Plot 2 - Large transparent spheres with individual RGB color values corresponding to their XYZ position.

Figure 12: Illustrates plots 1 and 2 from the first iteration design.



(a) Plot 3 - Small cubes, each of which is one of four different colors.



(b) Plot 4 - Large opaque spheres, with individual RGB color values corresponding to their XYZ position. This plot also includes a transparent plane

Figure 13: Illustrates plots 3 and 4 from the first iteration design.

These different dataset plot representations seek to explore as wide a base as possible so that we would be able to explore test participants' experience with, not only different colors, sizes, transparencies, shapes, and amounts of individual data points, but also vertical lines and an angled plane to explore how well users' interpret certain limitations and how well the AR data visualization affords viewing from different angles. This should hopefully give some broad feedback for

the general visibility of each of the data plots, not enough to figure out which is the most optimal, but instead to figure out if any of them are sub-optimal or worse.

We do not aim to test usability on a very large sample size (n = 8), so while we do not intend to find tendencies or conclude anything specific from these data plots, the intention is exclusively to have a more broad exploratory view and general feedback, which we then can use to iterate for future iterations, as well as find any glaring errors or problems that would need immediate fixing.

5.1.1.1 ARFoundation

The fundamental implementation of the AR functionality was the first step in the development of the first iteration prototype, as this was the essential functionality all other features would eventually rely on for better or worse. This meant that we would have to strive for optimal performance and compatibility from the get-go to avoid ending in a situation where we would have to discard all built-on-top functionality if the AR functionality would be subject to change during development. Luckily, the ARFoundation library, which had promising Unity and iOS integration and compatibility, would serve as a great and easy-to-implement package with a relatively high fidelity. Using ARFoundation's online documentation [18], the implementation process was relatively straightforward as the ARFoundation package contained a multitude of prebuilt functionality, with existing prefabs, scripts, and scene templates to explore.

The XR Origin and AR Session GameObjects act as essential AR components, with the XR Origin GameObjects having additional components added, depending on the context of the AR application; in this case, Plane detection and Raycast functionality (see Figure 14).

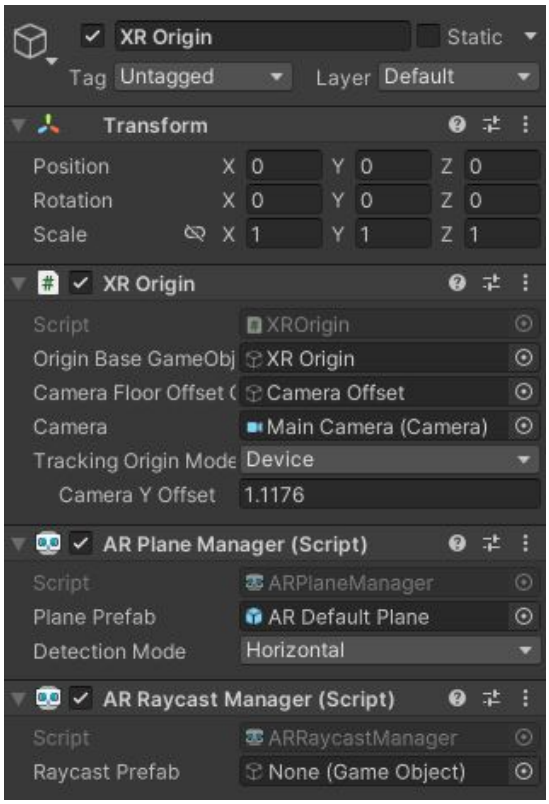


Figure 14: The XR Origin GameObject with ARFoundation components; XR Origin, AR Plane Manager, and AR Raycast Manager.

Noteworthy settings include the plane prefab assigned as the AR Default Plane and the detection mode set to Horizontal. The AR Default Plane is a plane prefab from the ARFoundation package that dynamically changes depending on the detected plane. In this case, the plane is instantiated during the plane detection and displayed, with some custom modifications to help visualize when, how much, and where a plane is being detected. This is useful not only for debugging but also functions as feedback about the plane detection to the user, which alternatively would be hard to know or understand. The detection mode, which is set to horizontal, helps the image recognition by only detecting perfectly horizontal planes. This ensures that the application does not accidentally detect a slanted floor by mistake, which could influence the rest of the AR experience negatively in a multitude of different ways.

As part of the development of the application being built on the iOS platform, testing and debugging specific AR-related aspects required attaching an iOS device and targeting the iOS platform when building through Unity, before finally launching and running the application in real-time. This was a slow and time-consuming process, however, it was a necessary step when developing to iOS and has resulted in a relatively consistent and efficient result of AR with good potential for affording motion tracking to navigate the augmented environment, as well as affording switching between portrait and landscape mode fluently.

5.1.1.2 Placement

To maximize compatibility with most given environments a user might use the application in and afford a level of customizability and ensure congruity with any given dataset, which might have differently scaled dimensionality, we wanted to afford dynamic plot placement functionality. That meant that we would have to develop an intricate placement feature as opposed to just having a fixed-scale plot placed at a certain location. For this feature, we figured a simple three-step process would be the most optimal and intuitive approach. The three steps in question were as follows (see Figure 15);

1. **Place point 1:** Places the first point, which will serve as one of the bottom corners of the entire plot. At this point, the rotation is also set according to the rotation of the indicator. The point is continuously calculated and updated, as feedback to the user, depending on the camera angle ray cast hitting the detected plane.
2. **Place point 2:** Places the second point, dragging out a rectangle shape between itself and the original point. This can go in either direction. The point is continuously calculated and updated, as feedback to the user, depending on the camera angle ray cast hitting the detected plane.
3. **Set height:** Set the height, using the rectangle as a base with a minimum height threshold. The height is continuously calculated and updated, as feedback to the user, utilizing a simple mathematical equation based on the vertical angle of the camera.

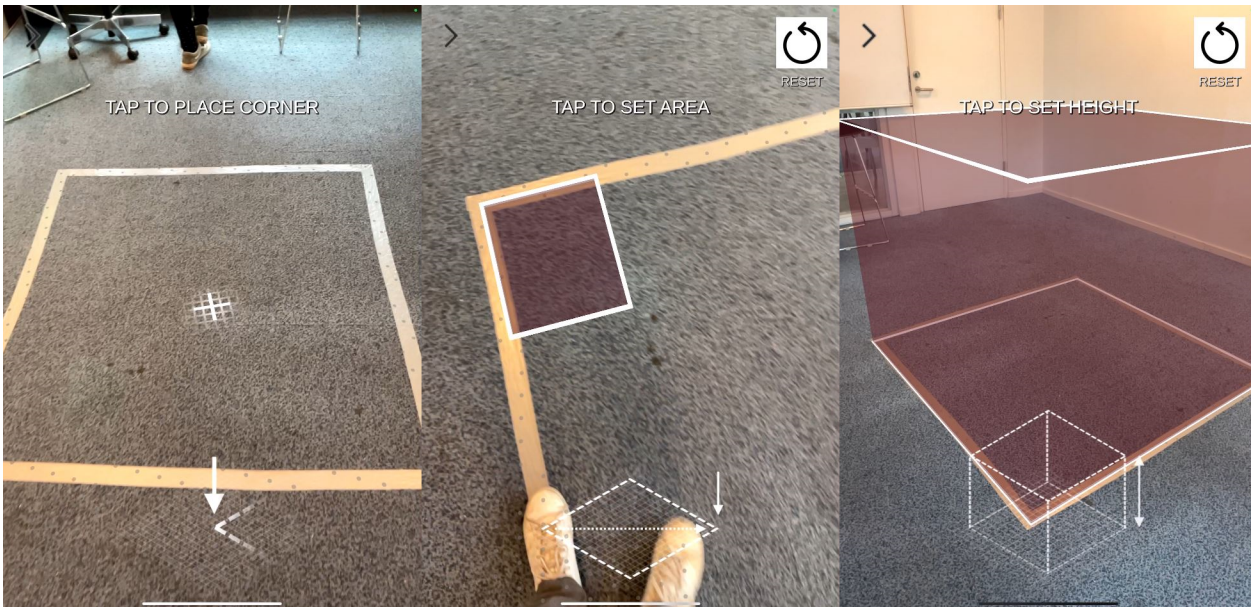


Figure 15: The three-step process for placing the plot in the AR environment, and their respective description text and icon imagery to guide the user. Duct tape was placed on the floor and acted as a guide when placing the area in the second step. Prior plane detection step not shown.

As shown in Figure 15, we have also added guiding text and icon imagery to best as possible help a user through the placement process as this can be an initially hard hurdle to understand intuitively.

```
1        if (raycastManager.Raycast(ray, hitList, TrackableType.PlaneWithinBounds))
2         {
3             findFloorPrompt.SetActive(false);
4             //find top plane
5             Pose hitPose = hitList[0].pose;
6             //update position
7             transform.position = new Vector3(hitPose.position.x, hitPose.position.y,
                  hitPose.position.z);
8             //update target and self position to align rotation of the indicator to the
                  camera direction
9             Vector3 targetPos = new Vector3(cameraObject.transform.position.x,
                  transform.position.y, cameraObject.transform.position.z);
10            Vector3 targetDirection = targetPos - transform.position;
11            transform.rotation = Quaternion.Euler(transform.rotation.x,
                  -cameraObject.transform.rotation.y*180, transform.rotation.z);
```

Figure 16: A segment in the Update() function of IndicatorScript.cs, which is placed on the indicator GameObject, continuously updates the position and rotation of the indicator based on the direction of the camera and a raycast hitting a plane prefab.

IndicatorScript.cs is a single script that controls the entirety of the placement stage of the application. Along with Figure 16, IndicatorScript.cs also contained sequential steps and Functions in the placement process, which would enable and disable specific GameObjects depending on the current step, using Unity UI Button elements that also would enable and disable each other. This would finalize the three-step process of placement, along with a reset button that would revert back to the first step in case the placement went wrong or the user wanted to do the placement differently.

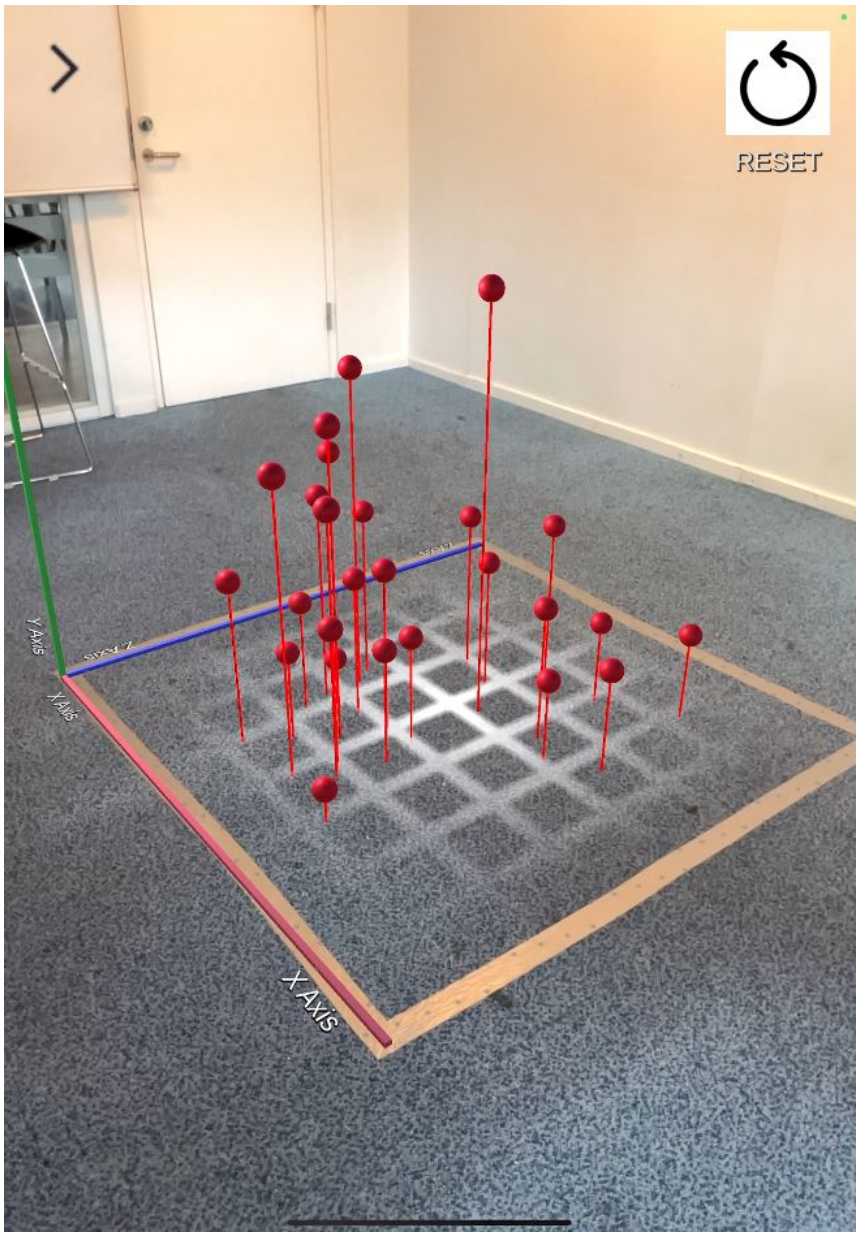


Figure 17: The plot placement after the placement process, showing dataset 1, augmented into the real-world environment.

After placement of the plot area, the next step is how the data is plotted in this dynamically shaped area, and how multiple datasets are switched between.

5.1.1.3 Dynamic 3D Plotting

The 3D plotting functionality is based on the placement of the plot and specifically uses information about where the three points are located and rotated in the 3D space. The PlaceCube() function in Figure 18 shows how the cubeObject, which acts as a parent object of the dataset plots,

is positioned, rotated, and scaled on each axis.

```
1     public void PlaceCube()
2     {
3         //enables the plot object and disables the indicator object
4         cubeObject.SetActive(true);
5         cubeIndicator.SetActive(false);
6         //set the correct position, rotation, and scale for the plot
7         cubeObject.transform.localPosition = new
          ↪  Vector3(point2.transform.localPosition.x/2, 0,
          ↪  point2.transform.localPosition.z/2);
8         cubeObject.transform.rotation = point1.transform.rotation;
9         cubeObject.transform.localScale = new
          ↪  Vector3(Mathf.Abs(point2.transform.localPosition.x/10),
          ↪  Mathf.Abs(point3.transform.localPosition.y/10),
          ↪  Mathf.Abs(point2.transform.localPosition.z/10));
10        //disables the indicator(itself)
11        gameObject.SetActive(false);
12    }
```

Figure 18: The PlaceCube() function, inside IndicatorScript.cs, used to finalize the placement of the data plot in the AR environment.

The transform of the plot, referenced as cubeObject in the IndicatorScript.cs script, uses localPosition as opposed to global position as the world space varies and is determined ultimately by where the user places the first point. Figure 19 shows the Unity hierarchy, and the parent/child relationship between the different points and the plot space, which is referred to as the DataCanvas.

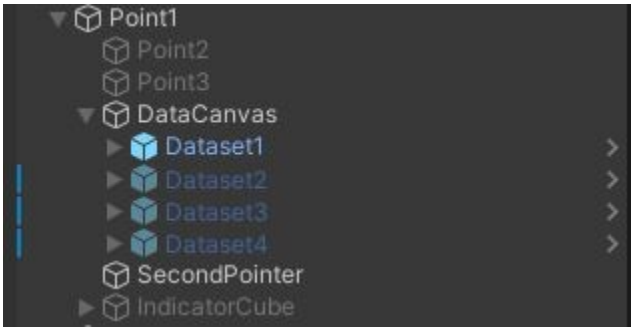


Figure 19: The Unity Hierarchy of the data plot showing the parent/child relationship between the Axis points and data plot called "DataCanvas" which has the actual 3D datasets as child GameObjects.

This hierarchy of GameObjects shows that Point 1, which is the initial first placement, determines

the rest of the GameObjects' local world space. This is useful as the following points; area and height have their local space relative to the original point allowing the entire placement and all the relevant GameObjects to inherit the translation and rotation offset from the otherwise arbitrary world space origin point. Additionally, the DataCanvas GameObject, once placed, functions as the new local space with unit scale measurements. This means any uniform prefab will be scaled according to the placed DataCanvas, which may have varying depth, length, or height so that it will fit at runtime.

The data plot prefabs (sometimes referred to as "datasets") are scaled as X=1, Y=1, Z=1 cubes so that in the future data points can be plotted with ease in the 3D space ranging from 0 to 1. This does however mean that the data plot prefabs will stretch in accordance to the varying placement. As the data point graphics should appear properly scaled independently of the plot, we added a scale constraint script called FixedScaleScript.cs (see Figure 20).

```
1   void Rescale(Transform obj, Vector3 newScale)
2   {
3       //ensures that it has a parent GameObject so returning the parent does not
        ↪ become redundant or throw an error
4       if(obj.root != obj)
5       {
6           //save the parent set the new parent to null
7           Transform parent = obj.parent;
8           obj.SetParent(null);
9           //scale it locally independant of any parent object and then set the parent
            ↪ object back so it still has the correct position and rotation
10          obj.localScale = newScale;
11          obj.SetParent(parent, true);
12      }
13  }
```

Figure 20: The Rescale() function, inside FixedScaleScript.cs, is used to scale the data points independently of the scale of the data plot. This ensures that uniform spheres or cubes will keep the correct scale as opposed to stretch in accordance with a stretched data plot. The position and rotation remain relative to the parent object.

This script is added as a component of every individual data point in the dataset prefab and runs the Rescale() function. This checks and disconnects the data point from the parent, and sets the localScale to a preset value before setting the data point back as a child GameObject of the dataset prefab. Setting the data point back as a child is important because we still want the data point to inherit the position and rotation if e.g. the plot would be reset and placed again. This also ensures that switching between datasets disables and enables them accordingly.

Lastly, the functionality of switching between different datasets, relies on all datasets being loaded

prior and disabled, except for the first dataset. A pop-out menu, which can be accessed in the top left corner (as seen in Figure 17), then has a button for each dataset that respectively disables all the datasets except the enumerated dataset which it then enables. This affords easy switching between datasets at any point and adheres to keeping the position and rotation consistent between all datasets.

5.1.2 Evaluation

The first iteration evaluation section will cover the testing setup and procedure while finally presenting the results of the usability test.

5.1.2.1 Testing Procedure

To prepare for the initial usability test the most recent and relevant studies on performing usability tests were analyzed (see Section 2.3.2.6). According to Albert and Tullis, a minimum of eight test participants are required and a proper testing environment must be established [48]. It was therefore decided to reach the minimum required amount, a total of eight test participants for the test. A closed-off medium-sized meeting room at Aalborg University Copenhagen was used as the testing environment. The flooring was fitted with a carpet and had a non-glossy and rough texture which provided an optimal surface for the AR prototype. The participants were from the Technical Faculty of Aalborg University in Copenhagen. Mostly stemming of 2nd and 4th semesters students and were as a result, not particularly experienced in the field of data visualization and analysis. The participants were asked to complete six different AR tasks which were ranked between the difficulties: easy, medium, and hard. After completing the tasks they would fill in the System Usability Scale (SUS) questionnaire and then answer six qualitative prototype-related questions specifically tailored by the researchers. The following list provides a detailed breakdown of the test procedure steps:

1. Sign the consent form.
2. The participant was led into the meeting room, introduced to the project, and encouraged to think-aloud while performing the tasks.
3. The participant completed Tasks 1 through 6, potentially interacting with the test facilitator.
4. The participant would fill in the 10-item System Usability Scale questionnaire.
5. The participant would respond to six qualitative and quantitative questions aimed specifically at the prototype solution.

The six tasks that the test participants were asked to complete were ranked in terms of difficulty: easy, medium, and hard. These estimations, were primarily to manage our expectations of the participants performance, but would for later iterations be updated continuously based on feedback and serve as a tool to estimate difficulty of different factors within tasks, and eventually present

tasks in order of easy to hard to compensate for a learning curve. The following table details the individual tasks they had to complete in accurate order:

Task	Task Difficulty	How to Pass
Scan the floor by moving around	Easy	They move around pointing the camera at the floor at different angles until the floor is detected
Place the area and set the height of the data plot. Try to match the area with the marked spot on the floor	Hard	They place the indicator by pointing and clicking followed by pointing and clicking at another area to form the square. They then adjust the height of the area by angling the camera up and down
Count the amount of data points	Easy	They must move around and count the data points. The correct amount of data points is: 26
Load dataset 2 and count the amount of data points	Medium	They must change to dataset 2, and count the data points. The correct amount of data points is: 29
Load dataset 3 and count the amount of red and green data points	Easy	They must change to dataset 3, and count the red and green data points. The correct amount of data points is: 22 (13 green+9 red)
Load dataset 4 and count the amount of data points above the plane	Medium	They must change to dataset 4, and count the amount of data points above the plane. The correct amount of data points is: 12

Table 2: This table gives an overview of the tasks, their difficulty level, and how to get them right.

Below are the individual questions from the six-item prototype-related questionnaire created by the researchers (see 10.4 Appendix D):

1. What level would you rate your statistical knowledge?
2. Did you prefer to use the Landscape or Portrait mode when visualizing the data?
3. Which dataset(s) was/were the easiest to interpret visually?
4. What made the chosen dataset(s) easy to interpret visually?
5. Which dataset(s) was/were the hardest to interpret visually?

6. What made the chosen dataset(s) hard to interpret visually?

5.1.2.2 Test Results

This section will present the results from the six tasks and the SUS questionnaire. It will also highlight some of the qualitative and quantitative responses gathered during the test, along with notes taken by the test observer.

In Table 3 you will find the results of the System Usability Scale questionnaire. The individual scores from each participant can be seen while the last row exclusively contains the SUS score, the composite score for the test.

Participant	SUS Score
1	75/100
2	70/100
3	92.5/100
4	67.5/100
5	72.5/100
6	55/100
7	87.5/100
8	85/100
Aggregated SUS Score: 75.625	

Table 3: The SUS scores based on the gathered data in the first iteration.

As seen in Table 3, the eight participants achieved a combined SUS score of 75.625 out of 100. To complement the SUS questionnaire results, qualitative and quantitative feedback was collected through six prototype-related questions. The first question asked the participants to rate their statistical knowledge in which they had to choose between three options: fundamental, intermediate, or advanced (see Table 1). five of the test participants chose fundamental while the remaining three rated themselves as intermediate. In the second question, they were asked whether they preferred portrait or landscape mode for visualizing the data. Five reported that they were indifferent, two preferred landscape mode and one preferred portrait mode. The participants were then asked to evaluate the difficulty of interpreting each dataset, as presented by the third and fifth questions. Five out of eight participants mentioned that Dataset 1 was the easiest to interpret, whereas the majority found Dataset 2 and 3 the most challenging. In the fourth and sixth questions, the participants were asked to describe which factors made the data interpretation easier. It was mentioned several times that the occlusion of data points reduced their ability to interpret the datasets correctly. On the other hand, it was noted that the vertical lines in Dataset 1 helped in distinguishing occluded data points and assessing the height of data points. However, the transparency of data

points resulted in confusion for some of the participants. One participant felt that the datasets were too clustered, while a few others remarked that the data points were too small, in particular, Dataset 3.

The usability test included six different tasks the participants had to complete, four of which were related to the datasets. In the four dataset tasks they had to count the number of data points present in the plot (see Figures 12 and 13). Table 4 provides an overview of the participants' task performance:

Participant	Task 1	Task 2	Task 3	Task 4	Knowledge Level
1	26/26	29/29	21/22	13/12	Fund. Lvl.
2	26/26	31/26	22/22	12/12	Interm. Lvl.
3	26/26	29/29	22/22	12/12	Fund. Lvl.
4	26/26	29/29	22/22	13/12	Fund. Lvl.
5	26/26	29/29	22/22	7/12	Fund. Lvl.
6	26/26	28/26	22/22	12/12	Interm. Lvl.
7	24/26	29/29	22/22	11/12	Interm. Lvl.
8	26/26	29/29	22/22	12/12	Fund. Lvl.
Correct Answers:	**7 out of 8**	**6 out of 8**	**7 out of 8**	**4 out of 8**	**Most Declared: Knowledge Level**
Binary Success:	**.87%**	**.75%**	**.87%**	**.50%**	**Fund. Lvl.**

Table 4: For each task, the participants had to count the number of data points based on certain conditions (see Table 2 for a description of each task.). The correct answers are colored in green while incorrect answers are colored in red. The binary success is calculated by giving correct answers the numerical value 1 and incorrect. Averaging turns it into percent. Fund. Lvl = fundamental level. Interm. Lvl. = intermediate level.

Table 4 clearly shows that Tasks 1 and 3 were the simplest, each having only a single incorrect answer. On the other hand, Task 2 resulted in two incorrect responses, and Task 4 was identified as the most challenging task with a total of four incorrect answers.

5.1.3 Discussion

The first iteration of our AR application was dedicated to examining its usability and user experience to establish a simple and stable foundation for the application's future additions. Here we will discuss some of the design decisions taken as well as the insights gained from the test results.

The testing environment was conveniently placed in near proximity of the Technical Faculty of Aalborg University in Copenhagen, which allowed for easy access to test participants. The testing environment, or meeting room, offered a secluded and interruption-free space only accessible by the researchers. Furthermore, the room had a carpet floor which provided a great texture for

the application's plane detection functionality. The meeting room used for the test was generally large in size, which means that there should be ample room for participants to move freely. If instances of participants attempting to use more extreme angles occur, and they feel restricted by the amount of space, it would be necessary to find a larger testing environment.

The datasets used in this iteration were imaginative data without any real meaning. The datasets and their respective visualizations were primarily aimed at testing the visibility and interpretability of data in a mobile-based AR solution. This was important in order to assess how well different visual cues such as data points translate into a 3D augmented space. It furthermore provided some insights into visual cues such as transparency color or height lines. As mentioned in the responses to the qualitative questionnaire (see 10.4 Appendix D), occlusion was an issue for some of the test participants and it could therefore make sense to include the options to enable/disable visual cues such as color or transparency in future iterations. Observations made during the testing showed that the participants found the application's user interface simple and intuitive. The participants found it easy to dynamically place and scale the data plot on the surface. This ease of interaction was significantly enhanced by the implementation of tooltips and pop-ups that guided the users through the process of placing and scaling the plot.

The usability test resulted in a System Usability Scale (SUS) score of 75.625/100, indicating a good level of usability. A score of 68/100 is declared as above average in the System Usability Scale. For a first iteration, these results are promising and can hopefully be improved in future iterations. It should be noted that the application still contains a very limited amount of features, which should be taken into account when discussing the SUS score. The results from the four dataset tasks each participant had to complete (see Table 4) showed that datasets with clear visual cues such as vertical lines in Plot 1 (see Figure 12a), were easier for the users to interpret. This might suggest that enhancing depth perception in AR visualizations could be used as a tool for improving understanding of the data. A general tendency for both the qualitative data and task data is that the responses were mixed. For example, the mixed feedback on landscape versus portrait visualization choices and the varying degrees of difficulty with different datasets that each participant experienced, proves the need for a more configurable and versatile design. This could be achieved by implementing more user control and customization in future iterations of development. During the task performance test, the participants were encouraged to think-aloud which provided valuable insights. Commonly, the participants mentioned a lack of clear indication of when the floor was successfully detected, which resulted in some uncertainty. It could therefore make sense to include some visual/audio feedback during plane detection in future iterations. Although the dynamic placement feature was appreciated by most participants, there were some users who had to reset several times in order to correctly refine their plot positioning. For example, the participants would sometimes position and scale the plot too flat resulting in them having to reset and adjust the plot's height for an easier view of data points. Having participants crouching

and shifting from portrait to landscape mode were also some of the strategies employed to gain better perspectives, particularly when counting data points or checking details within a dense dataset. Finally, the think-aloud observations included suggestions for additional features, such as increasing detail or differentiating data points more clearly as the user moved closer.

5.2 The Second Iteration

Based on the usability test results from the first iteration, we decided to make some decisions related on aspects to improve, add, and focus less on. For this iteration it was important for us to use actual datasets with different variables as opposed to arbitrary data points placed relatively randomly. This would mean that the datasets would have real statistical relevance as well as many more potential data points, and should prove to be a more optimal real-world use-case for the application. We will also focus on giving participants statistical tasks to solve, with the given datasets, and use the SUS questionnaire to evaluate. This should help to reflect a more realistic use-case as well as the potential use-case for the final testing. Additionally, this section will cover the integration of Data Point Highlighting and UI additions.

5.2.1 Development

As the first iteration was fairly fundamental in terms of the actual data functionality, this would be the core focus of this iteration. This ensures that we get to establish usability directed towards interpretation of real datasets as opposed to more fundamental task of such as counting data points. For this iteration, tasks would be related more towards identifying correlations, identifying categories, and estimating average values.

5.2.1.1 Dataset Integration

The two datasets we would focus on for this iteration was the Iris Dataset about specific flower species with variables such as sepal and petal widths and lengths, as well as the Red Wine Dataset with variables such as Sulphates, pH, Alcohol, Quality, Residual sugar, etc.

The first step taken by the application is running the LoadIrisData() function from the Data.cs script, which runs the function in Start(), i.e. the first frame of the application (see Figure 21).

```
1   void LoadIrisdataset()
2   {
3       using (var reader = new
        ↪  StreamReader(Path.Combine(Application.streamingAssetsPath, "Iris.csv")))
4       using (var csv = new CsvReader(reader, CultureInfo.InvariantCulture))
5       {
6           // Do any configuration to `CsvReader` before creating CsvDataReader.
7           using (var dr = new CsvDataReader(csv))
8           {
9               iris_dt = new DataTable();
10              iris_dt.Columns.Add("SepalLengthCm", typeof(float));
11              ...
12
13              ...
14              iris_dt.Load(dr);
15
16              NormalizeColumn(iris_dt, "SepalLengthCm");
17              ...
```

Figure 21: Excerpts from the LoadIrisdataset() Function which runs in Start(). The following columns being added and loaded which are not shown: ID, Sepal Width, Petal Length, Petal Width, and Species. A similar Function; LoadWineQualitydataset() is used for the Red Wine Dataset.

The LoadIrisdataset() Function in Figure 21 firstly loads the csv file from a specific Streaming Assets Path. This path is essential when it comes to building the application for devices, especially iOS, as the file directories often are more inaccessible during runtime of a mobile/tablet application. Using a csv Reader, we then read the actual csv file content and add the contents to a data table which we use later. The content is manually adjusted in the function by adding the columns by specific strings matching the ones in the csv file. A similar function for the Red Wine Dataset was of course also added. The final step of the load scripts are running the necessary columns through a NormalizeColumn() function (see Figure 22).

```
    foreach (DataRow row in dtz.Rows)
        {
            float value = Convert.ToSingle(row[columnName]);
            if (value < minValue) minValue = value;
            if (value > maxValue) maxValue = value;
        }
        foreach (DataRow row in dtz.Rows)
        {
            float value = Convert.ToSingle(row[columnName]);
            float normalizedValue = (value - minValue) / (maxValue - minValue);
            row[newColumnName] = normalizedValue;
        }
```

Figure 22: The two main steps of normalizing the a column of data in a dataset in the NormalizeColumn() Function.

This NormalizeColumn() Function is used when we want to display the data points on the 3D grid of the Unit cube in the AR environment. The functionality simply iterates through all the entries of a column and assigns the minimum and maximum values of all the entries. Those Min/Max values are then used to interpolate every single value to a value between 0 and 1. Only columns using number values are normalized as opposed to string values such as "Species" or "Names".

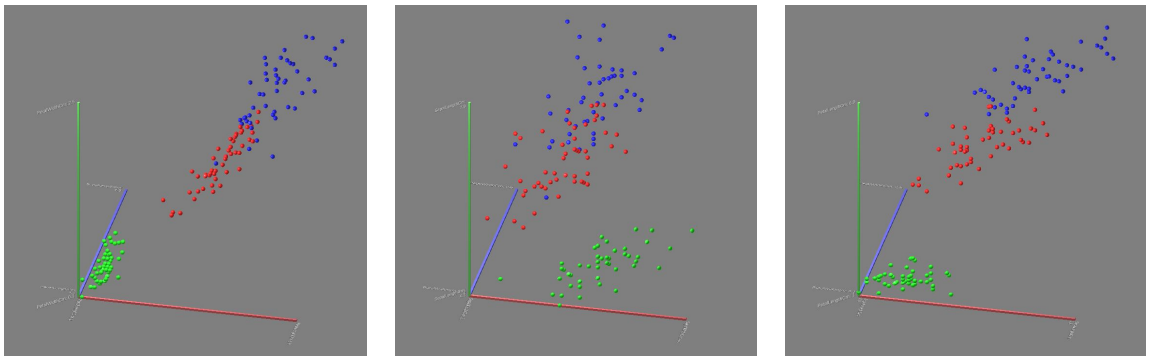
```
1   public void InstantiateIrisData1()
2   {
3       DestroyAllPrefabs();
4       EnableHelper(0);
5       if (iris_dt == null) return;
6       foreach (DataRow row in iris_dt.Rows)
7       {
8           GameObject prefab = Instantiate(dataPointPrefab, gameObject.transform);
9           prefab.transform.localPosition = new
            ↪ Vector3(Convert.ToSingle(row["PetalLengthCm_normalized"]),
            ↪ Convert.ToSingle(row["PetalWidthCm_normalized"]),
            ↪ Convert.ToSingle(row["SepalLengthCm_normalized"]));
10          prefab.GetComponent<DataPointScript>().dataDescriptionText = $"Sepal
            ↪ Length(cm): {row["SepalLengthCm"]}\nSepal Width(cm):
            ↪ {row["SepalWidthCm"]}\nPetal Length(cm): {row["PetalLengthCm"]}\nPetal
            ↪ Width(cm): {row["PetalWidthCm"]}\nSpecies: {row["Species"]}";
11          ...
12
13          ...
14          prefab.GetComponent<FixedScaleScript>().Rescale(0.02f);
15      }
16      AssignLabelText(xMin, xMax, iris_dt, "PetalLengthCm");
17      AssignLabelText(yMin, yMax, iris_dt, "PetalWidthCm");
18      AssignLabelText(zMin, zMax, iris_dt, "SepalLengthCm");
19  }
```

Figure 23: The InstantiateIrisData1() Function, showing the process of unloading loading specifically designed scatter plots using the pre-loaded datasets. In this case, the X,Y and Z axes will represent PetalLenthCm, PetalWidthCm, and SepalLengthCm respectively.

The final step of integrating the dataset comes in the Instantiate Functions; InstantiateIrisData**1/2/3**() and InstantiateWineData**1/2**() (see Figure 23). Each of these Functions are run based on interactions of the UI buttons that load the specific dataset after placing the plot in augmented reality. Apart from instantiating each data entry in 3D space from the normalized values, this function also adjusts some of the individual data points settings and variables. This includes the data description text used when highlighting data points, the scale of the object identical to the first iteration Fixed scale functionality, and color of the material depending on the "Species" variable of that data entry.

After iterating through the entire dataset, the final step is to set the preassigned 3D text labels that appear on the axes in 3D space, to match and show the Min and Max values of that axis. This should ensure that users have a better understanding of what variable is being represented on an

axis as well as the range of numbers the scatter plot falls under.

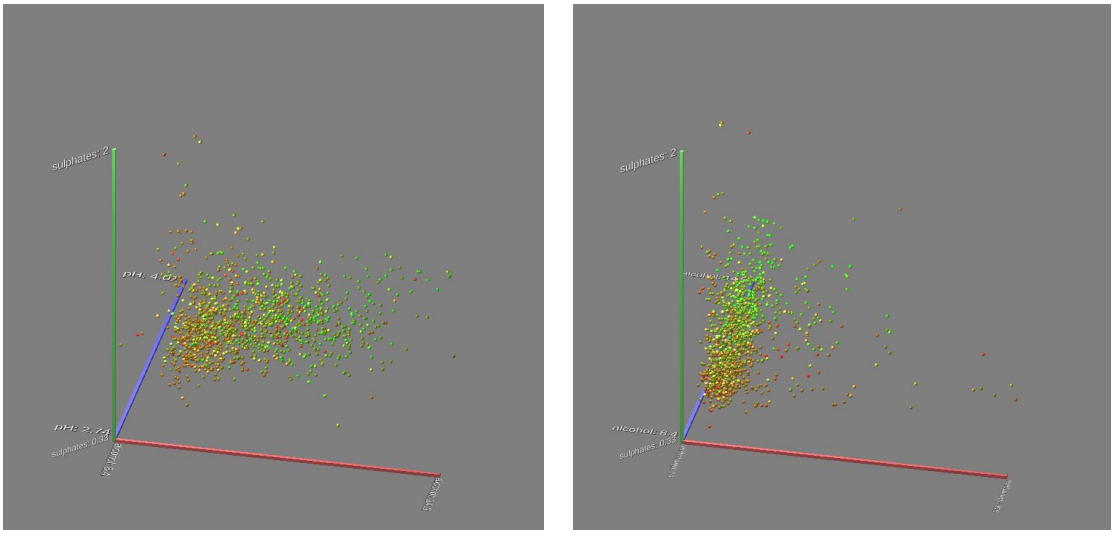


(a) Iris Plot 1 - Iris Dataset representing Petal Length(Red Axis), Petal Width(Green Axis) and Sepal Length(Blue Axis).

(b) Iris Plot 2 - Iris Dataset representing Sepal Width(Red Axis), Sepal Length(Green Axis) and Petal Width(Blue Axis).

(c) Iris Plot 3 - Iris Dataset representing Sepal Width(Red Axis), Petal Length(Green Axis) and Petal Width(Blue Axis).

Figure 24: Illustrates Iris Plots 1, 2 and 3 from the second iteration design. See Table 5 for respective tasks



(a) Wine Plot 1 - Wine dataset representing Alcohol (Red Axis), Sulphates (Green Axis) and pH(Blue Axis). The colors of individual data points also represent their Quality(Green/high - Red/low).

(b) Wine Plot 2 - Wine dataset representing Residual Sugar(Red Axis), Sulphates (Green Axis) and Alcohol (Blue Axis). The colors of individual data points also represent their Quality (Green/high - Red/low).

Figure 25: Illustrates Wine Plots 1 and 2 from the second iteration design. See Table 5 for respective tasks

These five Functions, which load different scatter plots from the Iris and Red Wine Dataset, are designed to have varying visualizations of the same data by using different variables for each of the three axes, as well as the color of the data points for an additional dimension. Specifically, these dataset scatter plots, were designed with the intention of having the possibility of different statistic related tasks/questions, which would require a test participant to have a more statistical approach when observing the datasets, as opposed to the first iteration which had a much more

fundamental counting approach, as mentioned previously (see Figures 24 and 25).

5.2.1.2 Data Point Features

As part of instantiating multiple entries in multiple columns of the datasets, the prefabs that are being instantiated also had some individual functionality. The Prefab that becomes instantiated for each data point has a child GameObject, with a mesh, material, and DataPointScript.cs attached. DataPointScript.cs allows the data point additional information to show or toggle between.

By detecting a screen tap, we attempt to distinguish if there is any data point being clicked by casting a ray from that point on the screen, which looks for a collider with a specific GameObject tag from the Prefab. If one is found that specific data point will enable a child-component which includes a highlight mesh, a line renderer, and 3D text label. If a ray is cast and no data point is hit, any previously highlighted data point will be disabled again.

```
1   void TapRayCast()
2   {
3       if(highlightedObject!=null)
        ↪ highlightedObject.transform.GetChild(0).gameObject.SetActive(false);
4       highlightedObject = null;
5
6       Ray ray = camera.ScreenPointToRay(Input.mousePosition);
7
8       if (Physics.Raycast(ray.origin, ray.direction, out RaycastHit hit) &&
        ↪ hit.transform.CompareTag("DataPoint")){
9           highlightedObject = hit.transform.gameObject;
10          HighlightObject();
11      }
12  }
```

Figure 26: The TapRayCast() Function in the TapInteractionScript.cs showing how the ray is cast, detects and highlights a hit object.

The TapRayCast() function in Figure 26, runs from a constant tap input check from the scripts Update() function, and includes a wipe of any previously highlighted objects to avoid users having to disable them manually. The HighlightObject() function not shown, has similar but inverted functionality as line 3 and 4 in Figure 26.

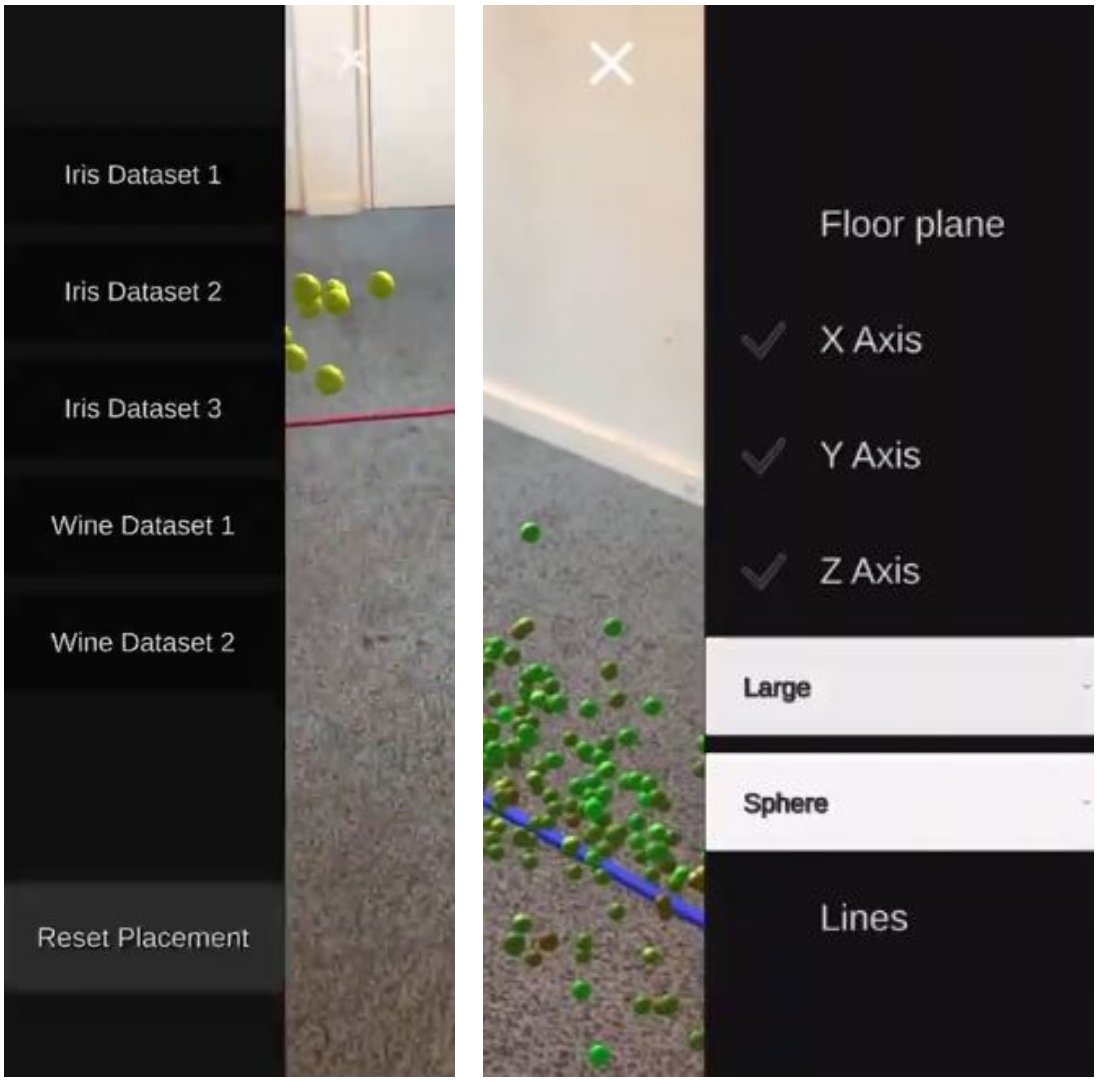
The text on the data point text label, includes multiple lines of information regarding the data points variables. This text label also includes a Billboard Script which constantly updates the

rotation on the Y-axis so that it always points to the camera. This avoids any situation where a user would not be able to read the text because of slanted or inverted text.

Lastly, a fade script was also implemented based on some of the feedback from the first iteration. This script essentially calculated the distance between all of the data points to the camera and linearly interpolated the transparency of the object based on a certain distance threshold. This caused a few issues in terms of performance on datasets with very high data point counts, as well as the fade material having issues rendering in correct order when occluded with various material and Line-renderers. As we also implemented a feature to scale the data points, we instead assume that the scale feature will solve the problem of data points being hard to differentiate between when clustered and/or occluded. This meant that we opted to not use this fade script feature for this iteration and future iterations going forward.

5.2.1.3 User Interface

As part of the improvements from the first iteration, we included a settings menu for additional features that could enable users to customize a few aspects such as data point size, shape, if they should have vertical lines, as well as some toggle-able plot visuals such as individual axes and the floor grid (see Figure 27b).



(a) The main menu in the top left showing the buttons that load each individual dataset as well as the reset placement button which allows the plot to be placed again.

(b) The settings menu in the top right showing the toggle options as well as the dropdown menus of data point size and shape.

Figure 27: Illustrates the Task and Settings menus, which the participants can access during runtime.

These menus were fitted with appropriate icon imagery and tapping outside of an open menu would close it to simplify the navigation. All the UI elements were built using Unity's built in UI system and button functionality, which allows us to run the toggle functions from the Data.cs script on button press.

```
    public void SetLine(bool toggle)
{
    var prefabList = GameObject.FindGameObjectsWithTag("DataPoint");
    foreach (var datapoint in prefabList){
        datapoint.GetComponent<DataPointScript>().SetLine(toggle);
    }
}

public void SetMesh()
{
    int mesh = meshdropdown.value;
    var prefabList = GameObject.FindGameObjectsWithTag("DataPoint");
    foreach (var datapoint in prefabList){
        datapoint.GetComponent<DataPointScript>().SetMesh(mesh);
    }
}
```

Figure 28: The SetLine() and SetMesh Functions in the Data.cs script which are toggled through the settings menu.

The SetLine and SetMesh() functions, as in Figure 28, allows us to loop through all the data points by searching the Unity hierarchy for Objects with the "DataPoint" tag, which exclusively includes the data point prefabs instantiated during runtime.

The final addition to the UI was a left-aligned label with text and color to help the user identify the meaning of data point colors in cases where relevant (see Figure 29).

Figure 29: The labels showing what each data point color represents; Species category(left) & Quality level(right)

This is an important feature to include as it would allow us to show an additional dimension for each data point using the color, which otherwise would be hard to interpret.

5.2.2 Evaluation

The following evaluation section will detail the testing setup and procedures for the second iteration test and conclude with a presentation of the usability test and task results.

5.2.2.1 Testing Procedure

The testing procedure in this iteration will focus more on assessing advanced features and settings related to specific data visualization tasks. The procedure mainly follows the same methodology as the first iteration usability test (see Section 5.1.2.1).

The second iteration testing was conducted in the same room used for the initial evaluation to reduce inconsistency in environmental variables (see Section 5.1.2.1). As previous, the testing setup included a test facilitator and a test observer. The facilitator introduced the application and its features to each participant, facilitating the testing process and handling all communications with the test participants. The test observer was responsible for noting the participant's responses to the tasks and recording any relevant observations. By observing participants interacting with the application, we can assess the user experience and uncover any UI or design flaws. Before the test, the participants would give their consent and were informed of their right to withdraw from the study at any time. After testing, the participants would have to rate their own familiarity with data visualization alongside the usability questions, to support the analysis of the results according to their knowledge and skills later on.

Participants were asked to complete a total of seven tasks, three concerning the Iris Dataset and four related to the Red Wine Dataset. The tasks had various difficulties, which in advance were categorized by the researchers as either easy, medium, or hard. The purpose of the tasks was to assess the application's ability to support complex data visualization and interpretation. The tasks themselves involved no time constraints, allowing participants to fully engage in the tasks. In the first iteration, the participants themselves were tasked with marking out the dimension of the scatter plot directly on the floor (see Section 5.1.2.1). This was included in order to assess the application's functionality in terms of plane detection and plot setting. However, for the second iteration, the focus is on evaluating new features and the application's ability to interpret 3D data. As a result, the participants were no longer required to place the plot themselves, in the second iteration test. Consequently, the duct-taped square previously used to guide participants on where to place the plot, was removed. In the second iteration, the researchers will be in charge of placing the plot before the actual test begins, ensuring that each participant views and interprets the same visualization with identical dimensions (width and height). The tasks themselves were explained both verbally by the facilitator and through written instructions posted on the wall, allowing the participants to refer to them when needed.

Task	Task Question	Task Difficulty	Correct Answer
Iris Dataset			
1	Determine which species generally has the longest petals	Easy	Iris-Virginica
2	Find the species with the narrowest sepal width	Easy	Iris-Versicolor
3	Examine the relationship between the length of the sepals and petals	Medium	Positive Correlation
Red Wine Dataset			
4	How does sulphate contents affect the quality of the wine	Easy	Neutral Correlation
5	How does alcohol contents affect the quality of the wine	Easy	Positive Correlation
6	How does residual sugar affect the quality of the wine	Medium	Neutral Correlation
7	Give an estimate of the average alcohol contents of all the wines	Hard	10-11%

Table 5: This table lists the specific tasks for each dataset, their difficulty level, and the correct answer.

After completing the tasks, the participants was asked to fill in the System Usability Scale (SUS) questionnaire and then respond to three quantitative statements designed by the researchers. These statements are specifically focused on the newly implemented features that allow users to click on individual data points to retrieve specific values. They were directly focused toward the prototype, providing more in-depth data about specific system features outside of the SUS questionnaire. The statements were as follows (see 10.5 Appendix E):

1. I found it intuitive to interact with the data points.
2. The data point information helped me to better understand the dataset.
3. I found the interface easily distinguishable within the AR environment.

5.2.2.2 Test Results

A total of eight test participants were involved in the second iteration test. As in the first iteration, the participants were mainly Medialogy students stemming of fourth and sixth semesters from the

Technical Faculty of Aalborg University in Copenhagen, and, as a result, slightly more experienced in the field of data visualization, compared to the participants in the first iteration (see Section 5.1.2.1).

The results from the second iteration test indicated a significant decline in usability scores. The System Usability Scale (SUS) score for the second iteration was 60 (see Table 6) which is less than the above average benchmark of the SUS framework (above average SUS is 68 or greater) and also lower compared to the first iteration's score of 75.625 (see Table 3).

Participant	SUS Score
1	77.5/100
2	52.5/100
3	80/100
4	27.5/100
5	75/100
6	55/100
7	57.5/100
8	55/100
Aggregated SUS Score: 60	

Table 6: The SUS scores based on the gathered data in the second iteration

As previously mentioned, the test participants were tasked with seven different assignments, three related to the Iris Dataset and four to the Red Wine Dataset. In Table 7, the participants' responses are color-coded for clarity: correct answers are indicated in green and incorrect answers in red. Additionally in Table 8, the categorization of task difficulty perceived by the participants' can be examined with their task difficulty score and the researchers assumed task difficulty.

	Iris Dataset			Red Wine Dataset				
Participant	**Task 1**	**Task 2**	**Task 3**	**Task 4**	**Task 5**	**Task 6**	**Task 7**	**3D Level**
1	Iris-Virginica	Iris-Versicolor	PosCorr	NeuCorr	PosCorr	NegCorr	10%	Not Fam.
2	Iris-Versicolor	Iris-Versicolor	PosCorr	NegCorr	PosCorr	NeuCorr	9.1%	Not Fam.
3	Iris-Virginica	Iris-Versicolor	PosCorr	NeuCorr	PosCorr	NeuCorr	10.4%	Very Fam.
4	Iris-Setosa	Iris-Versicolor	PosCorr	NegCorr	NegCorr	NeuCorr	9.0%	Not Fam.
5	Iris-Virginica	Iris-Versicolor	PosCorr	NeuCorr	PosCorr	NeuCorr	9.5%	Dece Fam.
6	Iris-Virginica	Iris-Setosa	PosCorr	NeuCorr	PosCorr	NegCorr	10.4%	Dece Fam.
7	Iris-Virginica	Iris-Setosa	PosCorr	NegCorr	PosCorr	NeuCorr	10.50%	Not Fam.
8	Iris-Virginica	Iris-Versicolor	PosCorr	NegCorr	PosCorr	NeuCorr	10%	Dece Fam.
Correct Answers:	**6 out of 8**	**6 out of 8**	**8 out of 8**	**4 out of 8**	**7 out of 8**	**6 out of 8**	**5 out of 8**	**Most Declared 3D Level:**
Binary Success:	**.75%**	**.75%**	**1.0%**	**.50%**	**.87%**	**.75%**	**62%**	**Dece Fam.**

Table 7: For each task, the participants had to provide the correct answer (see Table 5 for the individual tasks). The correct answers are colored in green while incorrect answers are colored in red. The binary success is calculated by giving correct answers the numerical value 1 and incorrect 0. Averaging turns it into percent. PosCorr = positive correlation. NeuCorr = neutral correlation. NegCorr = negative correlation. Not Fam. = not familiar. Dece Fam. = decently familiar. Very Fam. = very familiar.

	Iris Dataset			Red Wine Dataset			
Participant	**Task 1**	**Task 2**	**Task 3**	**Task 4**	**Task 5**	**Task 6**	**Task 7**
1	Medium	Medium	Medium	Easy	Easy	Easy	Medium
2	Medium	Medium	Hard	Medium	Easy	Hard	Medium
3	Easy	Easy	Easy	Hard	Medium	Medium	Medium
4	Medium	Medium	Hard	Hard	Medium	Hard	Easy
5	Medium	Easy	Easy	Medium	Easy	Easy	Easy
6	Medium	Medium	Hard	Medium	Hard	Medium	Medium
7	Hard	Easy	Easy	Medium	Easy	Easy	Easy
8	Easy	Easy	Medium	Medium	Hard	Hard	Hard
Assumed Task Difficulty:	**Easy**	**Easy**	**Medium**	**Easy**	**Easy**	**Medium**	**Hard**
Task Difficulty Score:	**43/100**	**25/100**	**50/100**	**56/100**	**38/100**	**50/100**	**38/100**
Most Declared Difficulty:	**Medium: 5 out of 8**	**Easy/Medium: 4 out of 8**	**Easy/Hard: 3 out of 8**	**Medium: 5 out of 8**	**Easy: 4 out of 8**	**Easy/Hard: 3 out of 8**	**Medium: 4 out of 8**

Table 8: Shows the the participants aggregated subjective rating of each tasks difficulty. The second to last row shows the difficulty level assumed by the researchers. The last row represents the calculated task difficulty score, which is calculated by averaging the participant ratings which are categorized as Easy (0), Medium (50), and Hard (100).

In the first task, the participants were required to identify the species with the longest petals. Six out of eight participants accurately answered that Iris-Virginica typically has the longest petals (see Table 7). Similarly, in the second task, the participants were asked to identify the species with the narrowest sepal width. Again, six out of eight participants correctly identified the species. After completing each task, the participants were asked to evaluate the difficulty level of the task they had just undertaken. For Task 1, the majority of the participants described the tasks as moderately challenging (see Table 8). This perception is contradictory to the researchers' initial classification of the task being easy, suggesting an underestimation of the tasks' complexity. In Task 3 within

the Iris Dataset, the participants had to analyze the relationship between sepal length and petal length, with all participants correctly recognizing a positive correlation. Although the researchers initially rated the task as a medium difficulty task, the participant's own assessment of the difficulty varied greatly, suggesting that previous experience with data visualizations and statistics might influence how challenging they find the task (see Tables 7 and 8).

Advancing to the Red Wine Dataset, Task 4 required the test participants to evaluate the impact of sulphate contents on wine quality, where only half of the participants answered correctly. When asked to assess the difficulty of the task, the participant responses consisted almost exclusively of medium ratings (five out of eight). This is contrary to what the researchers initially believed, assuming the difficulty was easy. In Task 5, they had to analyze the influence of alcohol contents on wine quality, where seven out of eight participants managed to answer correctly. Although the participants found this task easier than its predecessor, it still presented a greater challenge than initially anticipated by the researchers, reflected by a difficulty score of 38. However, the majority declared Task 5 to be of easy difficulty (see Table 8). Task 6 asked the participants to assess the effect of residual sugar on wine quality. Six out of eight participants correctly answered the task. The participants predominantly rated the task as easy or hard, with a composite score of 50, which aligns with the researchers' initial assumption of a medium task difficulty. The nature of the final task was different from the other six tasks as the participants had to estimate the average alcohol content of the wines. In order to answer this, the participants had to interact directly with the data points to make an informed guess. More than half of the participants were able to provide an accurate estimate within the required range of 10-11%. The task difficulty was perceived to be medium as half the participant declared this difficulty when asked, which stood in contrast to the researchers' prediction of a hard task.

5.2.3 Discussion

The second iteration delved into more complex design features, and tried to examine what features that would enhance the user experience and 3D data visualization. This included new and realistic datasets, features like: data point highlighting, axes labeling, and UI components.

The task results generally showed positive tendencies, although to some degree being contradicting in the assumed task difficulty and the actual task difficulty score. This could be due to the participants proficiency in 3D visualization as well as their statistical knowledge. Moreover, it could be due to their proficiency in using a unorthodox tool for 3D visualization, like our augmented reality application, compared to the usual choice of a 2D monitor display. Not being used to navigate a 3D environment with full AR integration might have resulted in a increasing task difficulty, and as a result we, the researchers, have misjudged the task difficulties. Nevertheless, the binary success for each task exceeds the 50% mark, indicating that although having difficulties in using the application the correct answer, for the most part, was found (see Table 7).

One thing is the task design and results, but we were also very interested in examining how the new features and changes to the already existing ones affected the usability of the application. As presented in Section 5.2.2.2, Table 6 can be found. When compared to the first iteration, which got a score of 75.625, the second iteration's usability has decreased. Although a SUS score of 60 is considered average, it is a great decline in usability compared to the previous iteration. This can be due to multiple aspects, with the major aspects being: new features and tasks. As we designed and implemented a multitude of new features, the complexity of the application raised. Although not groundbreaking features, they are more complex than the first iterations very limited features. This together with an increasing number of tasks, and more complex tasks than just counting, all together plays a role in the user experience. As a result, it is not unnatural to perceive a dip in performance, but we must process the data and adapt the application to conform to the user's needs. Adding more unique datasets and reworking the existing features to better accommodate the use-cases is expected to increase the System Usability Score. Also, acknowledging the task difficulty will support the overall user experience by aiding the users with improved features, namely by taking the difficulties of which tasks are hard into account, such as estimating values and identifying specific correlations being more difficult than initially expected.

5.3 The Third Iteration

Based on the Usability test results from the second iteration, we decided on a few areas of improvement, to get a final version of the prototype ready for the final experiment. Just like the previous iterations, this final iteration will include a usability test, this time using specifically designed tasks with varying difficulty based on the information we have gathered from the earlier iterations. The features that will be added for this iterations include adding additional datasets and dynamic axes. Importantly, am few features required for the final experiment will also be added, including tap gesture manipulation for the eventual control group and signal communication for the logging of task events.

5.3.1 Development

This iteration will focus on additional dataset integration with accompanying tasks and difficulty variation, gesture manipulation for the static version of the application, signal messaging from the application so we can track certain events from the tasks, and dynamic axes for easier navigation in the plot.

Despite this iteration does not test the physiological measures and tracking task performance, it is important that this functionality is set in stone for the final testing session and that we ensure it works without errors.

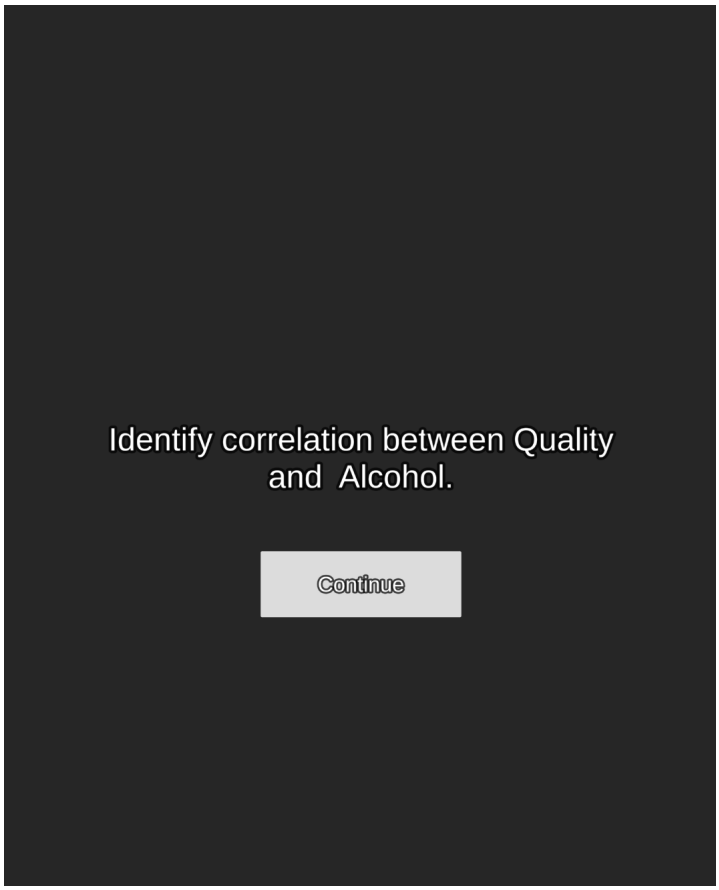
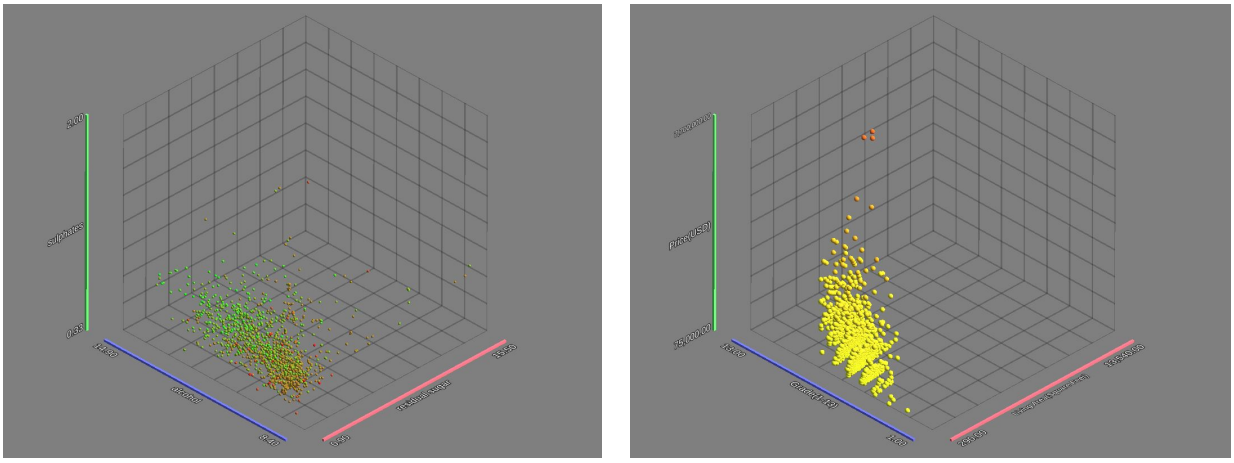


Figure 30: The Task presentation during runtime, with the Description of the task and a continue button which appears after a period of five seconds. The continue button will close the window and send an event signalling the beginning of the task. The task description will continue to display at the bottom of the screen so participants always can read it.

An example of the task implementation can be seen in Figure 30, which shows the final task prompt shown to a test participant after pressing *Task 1* in the main menu. This prompt is shown with a continue-button that only appears after a five second timer, to ensure that test participants do not immediately press the continue-button. This is essential as the continue-button closes the prompt window and starts data tracking the task (creates and logs events and tracks time) and the physiological measures. The actual prompt will be present during the task completion so test participants can read it again, if they want to. During a task, all Task-buttons in the main menu are disabled and a new button labelled *Finish Task* becomes enabled. This *Finish Task*-button is what we will tell participants to click on once they have answered the task question. This will enable the task-buttons again while data tracking the task. This will be covered more specifically in Section 5.3.1.3.

5.3.1.1 Additional Datasets

One of the significant changes from the second iteration to this iteration is the inclusion of three additional datasets: Housing, Movies, and Height and Weight (BMI) classification, combined into a total of five different datasets. These datasets would each be loaded and instantiated similar to how the Iris and Red Wine datasets were loaded and instantiated in the second iteration, with all of the compatible number variables being normalized as well.

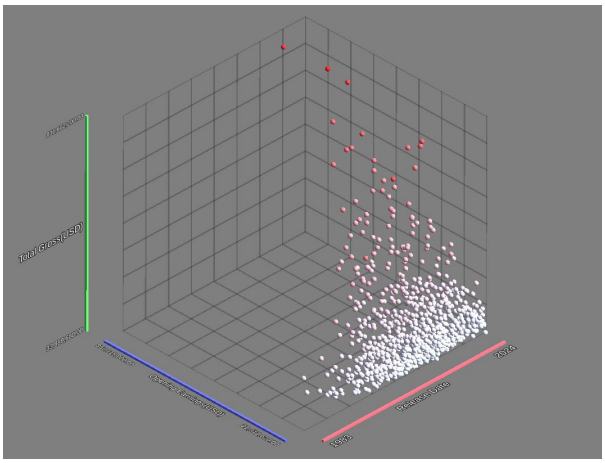


(a) Task 1 using the Wine dataset, with axes representingresidual sugar(X),sulphates(Y), and alcohol(Z). In this case the color of the data points represent Quality(green=high quality & red=low quality). The related task was formulated as follows: *"Identify correlation between Quality and Alcohol"* with the correct answer being *"Positive Correlation"*

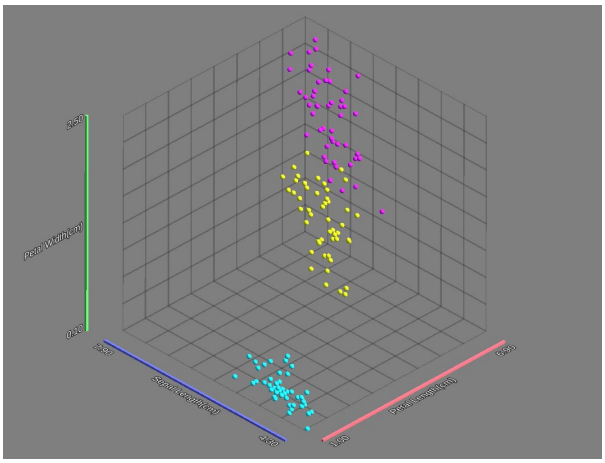
(b) Task 2 using the Housing dataset, with axes representingLiving Area in square feet(X),Price in USD(Y), and Grade(Z). In this case the color of the data points also represent Price(red=high price & yellow=low price). The related task was formulated as follows: *"Identify correlation between Living Area (sq ft) and Price"* with the correct answer being *"Positive Correlation"*

Figure 31: Two of the dataset plots; Red wine dataset(a) and Housing dataset(b)

The reason for choosing these specific datasets were that they would be relatively easy to interpret by test participants without external knowledge, as all the datasets are fairly fundamental as opposed to in-depth or niche ones. These datasets would also allow for us to formulate a variety of tasks related to different types of easy to interpret data visualization and statistic related questions. The final questions were based on the following: relationship between an axis and data point colors (Task 1 - Figure 31a), relationship between two axes (Task 2 - Figure 31b), finding a specific data point with given information (Task 3 - Figure 32a), defining a relationship between all three axes (Task 4 - Figure 32b), and estimating the position of a new data point with given variables based on the tendencies in the scatter plot (Task 5 - Figure 33).



(a) Task 3 using the Movies dataset, with axes representingRelease Date (X),Total Gross USD (Y), and Opening Earnings in USD (Z). In this case the color of the data points also represent Total Gross (red=high gross & white=low gross). The related task was formulated as follows: *"The Dark Knight has the highest total gross from any movie in 2008. Find and highlight the data point, and figure out what the exact release date was"* with the correct answer being *"18-07-2008"*



(b) Task 4 using the Iris dataset, with axes representingPetal Length (X),Petal Width (Y), and Sepal Length (Z). In this case the color of the data points also represent their respecive species(yellow=versicolor, cyan=setosa, & magenta=virginica). The related task was formulated as follows: *"Identify correlation between Petal Length, Sepal Length, and Petal Width"* with the correct answer being *"Positive Correlation"*

Figure 32: Two of the dataset plots; Movies dataset(a) and Iris dataset(b)

As opposed to Task 1 and 2, which we estimate to be relatively easy and serve more as introductory tasks, Tasks 3, 4, and 5 were estimated to be more difficult. A feature that was not included in the final design was the possibility of presenting numbers on an axis in intervals. This would make Task 3, and possibly Task 5, more easy to visualize, however, it was deemed out of scope as the importance of the difficulty in this case was not of a high priority since it would just mean that participants would have to rely on estimations. Task 4 was also expected to be slightly hard compared to Task 1 and 2 as we would be comparing three variables instead of two, however, the relationships between them were all pretty straightforward in the sense of all of them having a positive correlation, meaning it should prove to be passable to someone with limited statistical/data visualization knowledge/level.

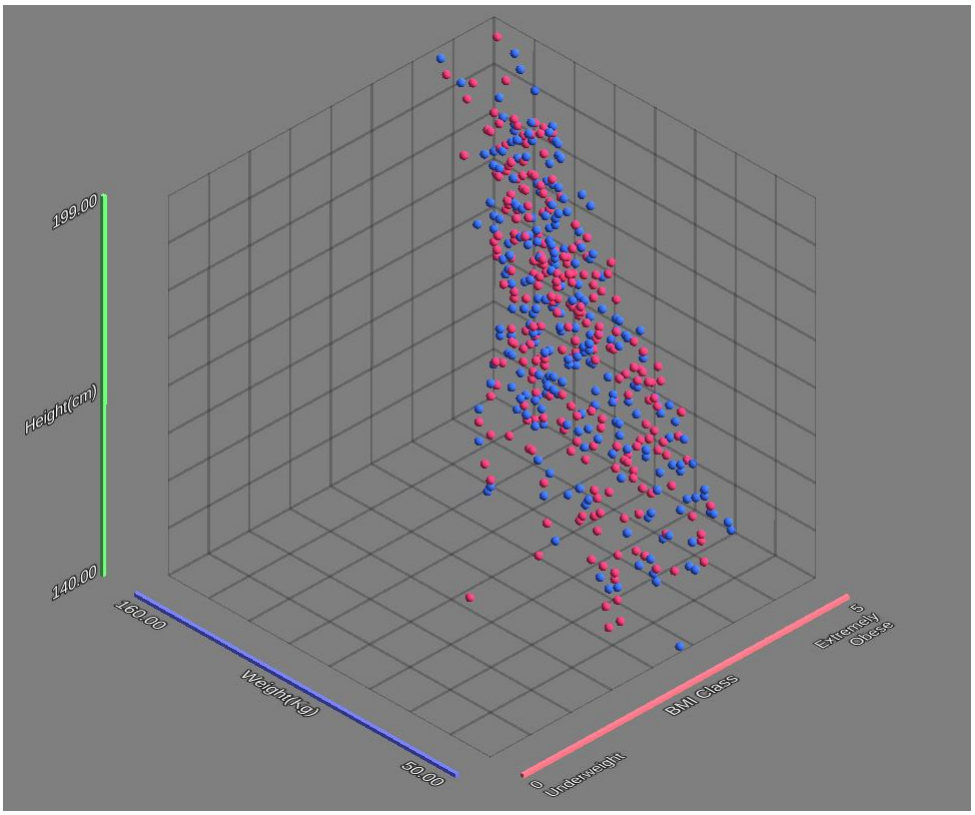


Figure 33: Task 5 using the BMI classification dataset, with axes representingBMI class(X),Height(Y), and Weight(Z). In this case the color of the data points represent their respective gender(blue=male & magenta=female). The related task was formulated as follows: *"Predict what BMI Category a person would have if they are 170 cm tall and weigh 100 kg"* with the correct answer being *"Category 4"*

The order of presenting the different tasks was based off difficulty as opposed to randomly. The idea behind this was that any disparity coming from the order of the tasks in the final test would be the same for both groups as long as the order of tasks being presented was the same. The priority was instead to present them in order of ascending difficulty to ensure all participants having an easy learning opportunity before heading into the harder tasks, which otherwise could have put them off and result in a much lower success rate across the board.

When it comes to loading a dataset, an important step that has not been discussed much in the second iteration is how the labels on the axes are applied.

```
1   private void AssignLabelText(TMP_Text minLabel, TMP_Text maxLabel,DataTable
    ↪ dataTable, string columnName, TMP_Text nameLabel, string nameOverride = null)
2   {
3       float minValue = float.MaxValue;
4       float maxValue = float.MinValue;
5
6       foreach (DataRow row in dataTable.Rows)
7       {
8           float value = Convert.ToSingle(row[columnName]);
9           if (value < minValue) minValue = value;
10          if (value > maxValue) maxValue = value;
11      }
12
13      minLabel.text = minValue.ToString("#,##0.00");
14      maxLabel.text = maxValue.ToString("#,##0.00");
15      nameLabel.text = nameOverride ?? columnName;
16  }
```

Figure 34: The AssignLabelText() Function which shows how the parameters are used when assigned during the instantiating of a dataset. The function is run for each individual axis, and includes the DataTable and columnName alongside the actual objects from the unity hierarchy assigned elsewhere. An additional Override string is also included which allows us to write a custom variable name that might be easier to interpret, e.g. changing "sqft_living" to "Living Area (Square Feet)". The function then iterates through all the values of the set columnName of the DataTable and establishes the new Min and Max values similarly to how they are calculated in the normalization process. Finally the label texts are assigned their new respective values with specific string parameters ensuring proper decimal format, e.g. changing 2785971 to 2,785,971.00

In Figure 34, the AssignLabelText() function uses multiple parameters. Previously in the second iteration these axis labels were only applied as minimum and maximum values across the three axes. For this iteration we not only expand the amount of axes by implementing new dynamic axes, which will be covered in Section 5.3.1.4, but also optimize the readability by adding a new 3D text label in the middle, allowing the Min/Max labels to only show numbers as seen in Figure 39.

5.3.1.2 Tap Gesture Manipulation

For the control group in the final test, i.e. the static version, we needed a way for the user to manipulate the plot without moving around in the AR environment. For this we opted to implement an external package from the Unity asset store called Lean Touch [105], which provided great tools, resources, and documentation readily available [106]. This would also greatly save time in having to develop our own tap gesture manipulation scripts. Having to implement our own scripts could be a risk as we would not be able to refine and polish them in time for the final testing. This could

result in poor performance in the control group, which inherently should strive to be as close to state-of-the-art of tap gesture manipulation.

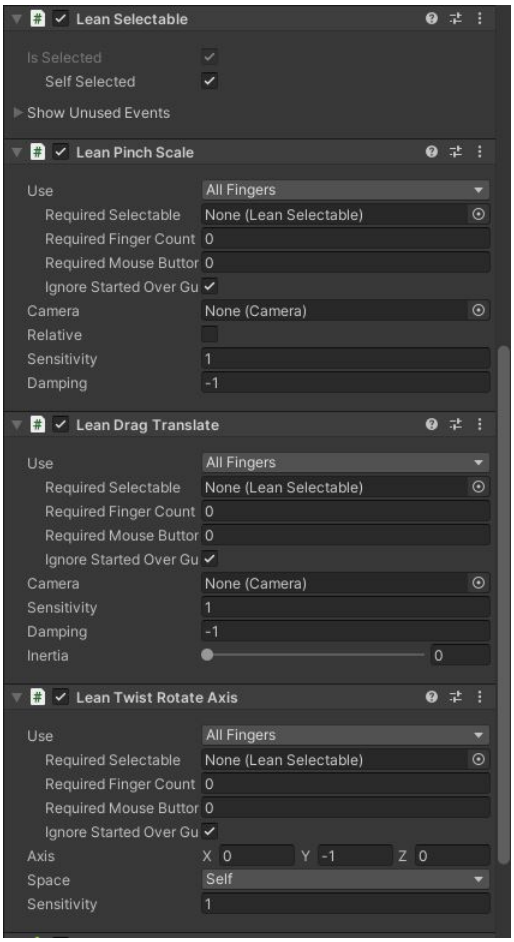


Figure 35: The four scipts from the Lean Touch package; Lean Selectable, Lean Pitch Scale, Lean Drag Translate, and Lean Twist Axis, all assigned to the plot GameObject. Here we can see the different parameters for each of the different interactions, including sensitivity, conditions, and orientations. A few of these empty required fields such as "Required Selectable" and Camera, are assigned during runtime automatically.

The implementation of the Lean Touch assets was fairly straightforward as they merely included assigning the specific scripts we wanted as manipulation to the actual 3D plot object as seen in Figure 35. In this case: Pitch Scale, Drag Translate, and Twist Rotate are familiar gestures, and are used in many touch-applications. Essentially, using one finger would move the position of the plot relatively to the viewing angle. Using two fingers would adjust the distance between them, resulting in scale functionality, and by twisting your two placed fingers you would rotate the plot around the Y-axis. A Lean Selectable script was also applied to ensure we mark the object as an interactable object in the Lean Touch scope, and an external GameObject had a global LeanTouch.cs script attached which is responsible for converting all mouse-and-tap input into more easily accessible data.

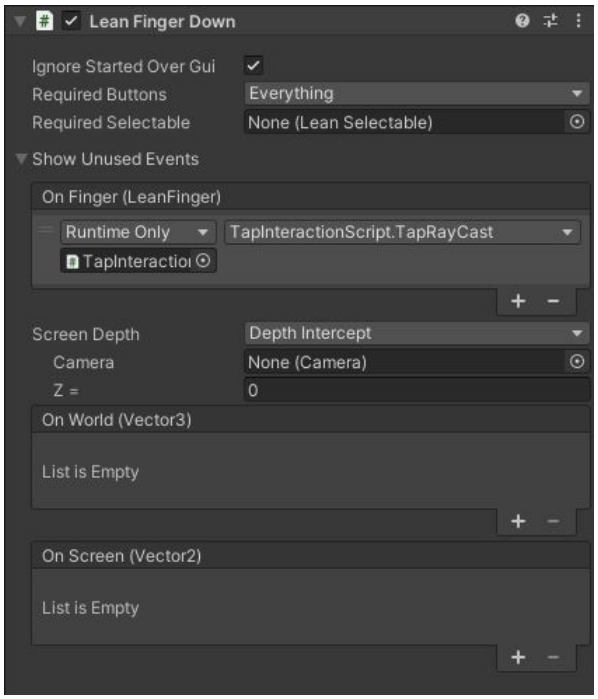


Figure 36: The Lean Finger Down script assigned to the TapInteractionManager GameObject, showing the different conditions for sending an event to the TapInteraction.cs script and running the TapRayCast() function. A few of these empty required fields such as "Required Selectable" and Camera, are assigned during runtime automatically.

The final Lean Touch script used was the Lean Finger Down script, (see Figure 36) which was assigned to the TapInteractionManager GameObject, had the previous used tap gesture highlighting component as well. This was an important step as implementing Lean Touch as a whole made the previous tap interaction non-functional, because of the input recognition. Instead Lean Finger Down recognizes a tap gesture in the Lean Touch scope and runs the old function, which has remained untouched, instead.

An important aspect to mention related to the Lean Touch implementation is that we prioritized moving the plot around using translate as opposed to rotating on the drag gesture. This could have allowed rotation on the other X and Z axes which could have made viewing the plot from "above" or at specific extreme angles much easier, however, this would have caused a larger disparity between the control and experimental groups than desired. We also valued being able to translate functionality higher in general because it would allow users to zoom in and out and focus more easily on areas that were not the center of the plot.

5.3.1.3 Signals

As mentioned previously, the following functionality would not be operational during the third iteration usability test, but rather be implemented to ensure that the application still runs perfectly fine without issues. The actual testing of the functionality would be done internally prior to the

usability test of this iteration and the pilot testing before the final experimental test, where all the necessary equipment would be equipped and event data tracked.

```
1   public void SendData(string key, string data)
2   {
3       string json = $"{{"{key}":"{data}"}}";
4       StartCoroutine(PutData($"{databaseUrl}messages.json?auth={apiKey}", json));
5   }
6
7   IEnumerator PutData(string url, string json)
8   {
9       var request = new UnityWebRequest(url, "PUT");
10      byte[] jsonToSend = new System.Text.UTF8Encoding().GetBytes(json);
11      request.uploadHandler = (UploadHandler)new UploadHandlerRaw(jsonToSend);
12      request.downloadHandler = (DownloadHandler)new DownloadHandlerBuffer();
13      request.SetRequestHeader("Content-Type", "application/json");
14
15      yield return request.SendWebRequest();
16
17      if (request.isNetworkError || request.isHttpError)
18      {
19          Debug.LogError("Error: " + request.error);
20      }
21      else
22      {
23          Debug.Log("Data sent successfully! Status Code: " + request.responseCode);
24      }
25  }
```

Figure 37: The The SendData() and PutData() functions in the MessageSystem.cs script, which uses the Firebase Realtime Database API to used to send messages at runtime from the Unity application for specific event markers. In this case we use a public SendData() function to start a coroutine of the PutData() method, which connects with Firebase API and inserts time markers for specific events.

Using this script, we connect to the Firebase Realtime Database with an API key to access the functionality of the Firebase database and send/receive messages across devices into the database (see Figure 37). The specific use case of this messaging system is to track the task time during the tests. We do this by sending messages labelled "Task X Started" (where "X" is the task number) and "Task Finished". These messages will be time stamped in the gathered data, so that we are then able to determine the time window in between starting and completing each task for each test participant.

5.3.1.4 Dynamic Axes

The final feature implementation covered in this development section is the dynamic axes feature. This feature was based on some of the feedback from the second iteration related to issues on locating and reading axes, especially when viewing from the other side. The solution was to create a dynamic system that continuously changes the position of the axes depending on where the user/camera is, relative to the plot.

```
1   void ActivateAdjacent(GameObject[] axisObjects, float minDistance)
2   {
3       int minIndex = -1;
4       float minDifference = float.MaxValue;
5
6       for (int i = 0; i < axisObjects.Length; i++)
7       {
8           float distance = Vector3.Distance(axisObjects[i].transform.position,
            ↪  transform.position);
9           float difference = Mathf.Abs(distance - minDistance);
10          if (difference < minDifference)
11          {
12              minDifference = difference;
13              minIndex = i;
14          }
15      }
16
17      int adjacentIndex = (minIndex + 1) % axisObjects.Length;
18      foreach (GameObject go in axisObjects)
19      {
20          go.SetActive(false);
21      }
22      axisObjects[adjacentIndex].SetActive(true);
23  }
```

Figure 38: The ActivateAdjacent() function used for determining which corner the green Y axis should be shown. All the distances between the camera position and each individual corner axis are calculated elsewhere, and the minimum distance, i.e. the distance of the closest axis, is used and compared as a parameter to find out which axis GameObject index is the corresponding axis. A modulo expression is used to loop around on line 17 to find the index next to the closest object, before setting all but the indexed GameObject to false.

Apart from the ActivateAdjacent() function in Figure 38, the X and Z axes use a simple Activate-Closest() function which is much more simplified in finding the closest of two GameObjects, and setting the closest to true. The visual appearance of this effect, along with an updated grid that

only appears on the inside of the plot, can be seen in Figure 39

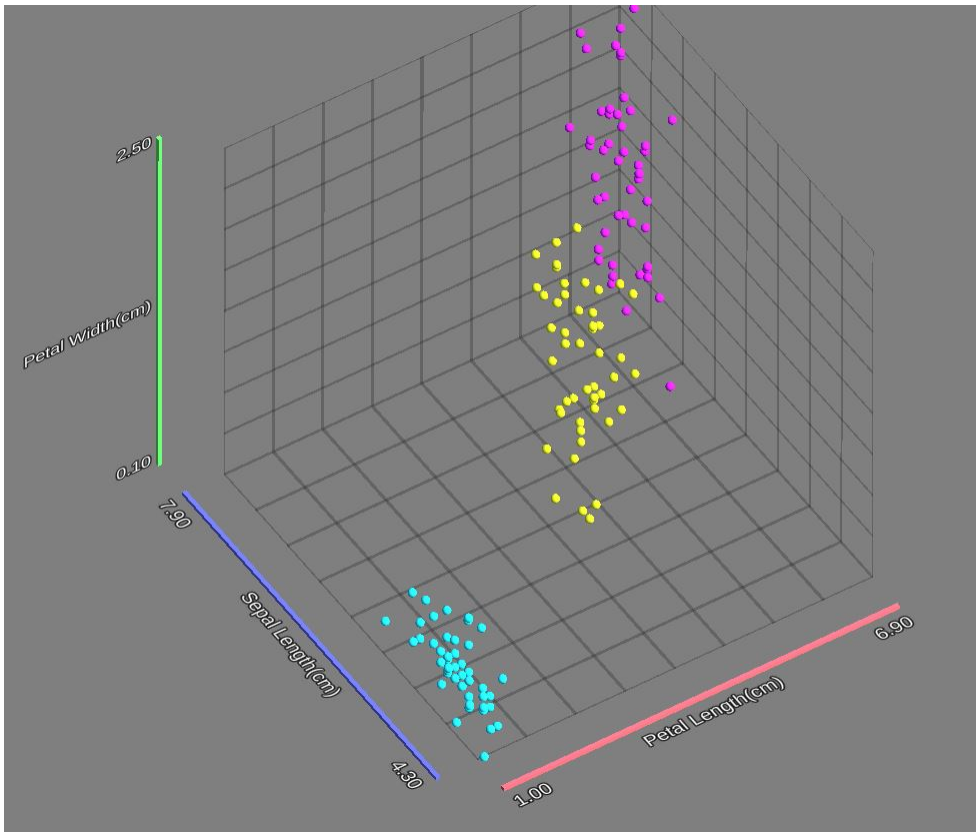


Figure 39: The plot as seen when the camera position is in the closest quadrant when viewing from this angle. Once the camera position is moved to another quadrant and is closer to a different corner than previous, the axes will switch accordingly. This ensures that the two bottom axes are always easy to navigate to and read, alongside the vertical axis which is always on the left hand side(as opposed to the directly closest corner which could occlude the scatter plot).

While this feature is expected to be a clear improvement to the previous static axes solution, the axes can still prove to be slightly confusing if the users position is equally far between two corners which could cause rapid switching between the active axes. This is especially confusing the closer to the middle you are, where the exact position might not match where the user expects the axes to be. This is a rather niche scenario that we do not expect to run in to that much during testing.

5.3.2 Evaluation

The following evaluation section will present the testing setup and procedure used in the third iteration test. Lastly, the results from the usability test and tasks will be presented.

5.3.2.1 Testing Procedure

This subsection will outline the testing procedure used for this iteration, which is very similar to the former iteration (see Section 5.2.2.1). The test will focus on general usability and testing the newly implemented UI features for task management, which essentially allows the users to read and initiate tasks via the UI rather than relying on communication via the test facilitator.

The third iteration test maintained the use of the same controlled environment that was used in the two previous tests, a medium-sized meeting room located at the Technical Faculty of Aalborg University in Copenhagen. Much like the previous iterations, this iteration recruited participants across the Technical faculty of Aalborg University, including primarily Medialogy students. Before beginning the test, the participant would have to rate their familiarity with data visualization into three levels. Participants were then briefed about the test, the AR application, and the tasks they would have to perform. The participants would then complete five tasks of varying difficulties (easy, medium, hard). In this iteration, each task was linked to a new and unique dataset, unlike the previous iteration where a single dataset was used for multiple tasks. Furthermore, the duration of each task was recorded to measure how long participants spent completing each task. This information will help us decide whether to implement task-specific time limits in the final test.

A significant change in this iteration was the introduction of new UI features for task interaction. Participants were required to start each task by clicking a 'continue button', but only after spending a mandatory five second duration where the participant would be presented with the task description on the iPad screen. This feature allows us to create accurate timestamps for each task's start and end, which is essential to analyze the physiological data later on. Therefore, the task details were no longer read aloud by the test facilitator but instead integrated within the AR application itself. After completing the tasks, the participants would fill in a System Usability Scale (SUS) questionnaire.

The tasks that a participant had to solve can be seen in Table 9. Presented are the five tasks, their estimated task difficulty, and the correct answers.

Task	Task Question	Task Difficulty	Correct Answer
Red Wine Dataset			
1	Identify correlation between Quality and Alcohol	Medium	Positive Correlation
Housing Dataset			
2	Identify correlation between Living Area (sq ft) and Price	Medium	Positive Correlation
Movie Dataset			
3	The Dark Knight has the highest total gross from any movie in 2008. Find and highlight the data point, and figure out what the exact release date was	Medium	18-07-2008
Iris Dataset			
4	Identify correlation between Petal Length, Sepal Length, and Petal Width	Medium	Positive Correlation
Height and Weight dataset			
5	Predict what BMI Category a person would have if they are 170 cm tall and weigh 100 kg	Hard	4

Table 9: This table lists the specific tasks for each dataset, their difficulty level, and the correct answer.

5.3.2.2 Test Results

For this usability test, a total of eight participants were involved, all of which were sampled from the Technical Faculty of Aalborg University, primarily Medialogy students.

For this iteration, improving the SUS was one of the main targets. The above average benchmark of the SUS score is 68/100. As seen in Table 10 three out of the eight participants scored the usability higher than the benchmark, with an additional two participants following close with a score of 67.5. However, due to a limited amount of participants (eight in total) there is little space for outliers. With two participants scoring the usability of the application 55 and 57.5, it greatly diminishes the chances of exceeding the 68/100 benchmark. Consequently, the aggregated SUS score for this iteration's usability was: 66.25/100. We concede that the application did not meet the 68/100 threshold we were aiming for, however we estimate that there are a multitude of reasons for this score being lower than anticipated, such as only having a sample size of n = 8, and that we are working with a fundamentally complex system. Considering this, we accept the current applications state of development and plan to continue using it for the final experimental test

Participant	SUS Score
1	55/100
2	57.5/100
3	72.5/100
4	75/100
5	75/100
6	67.5/100
7	67.5/100
8	60/100
Aggregated SUS Score: 66.25	

Table 10: The SUS scores based on the gathered data in the third iteration

Despite sampling from a similar population as earlier iterations, the participants for this iteration were less experienced with data visualization overall. This can be perceived when examining Table 11 where the most declared level of 3D visualization is *Not Familiar*. In the second iteration the level was *Decently Familiar* (see Table 7). Although not familiar with 3D visualizations, the majority of the participants managed to solve all of the tasks, with the least solved task being Task 4, having a total of three incorrect answers. In contrast, all participants managed to find the correct answer for Task 2. Task 3 and 5 each had two incorrect answers, while Task 1 only received one incorrect answer.

	Red Wine Dataset	Housing Dataset	Movie Dataset	Iris Dataset	HW Dataset	
Participant	**Task 1**	**Task 2**	**Task 3**	**Task 4**	**Task 5**	**3D Level**
1	PosCorr	PosCorr	Correct	Unsure	3	Not Fam.
2	NeuCorr	PosCorr	Unsure	NegCorr	4	Not Fam.
3	PosCorr	PosCorr	Unsure	PosCorr	3	Not Fam.
4	PosCorr	PosCorr	Correct	PosCorr	4	Not Fam.
5	PosCorr	PosCorr	Correct	PosCorr	4	Dece Fam.
6	PosCorr	PosCorr	Correct	PosCorr	4	Not Fam.
7	PosCorr	PosCorr	Correct	Unsure	4	Dece Fam.
8	PosCorr	PosCorr	Correct	PosCorr	4	Dece Fam.
Correct Answers:	**7 out of 8**	**8 out of 8**	**6 out of 8**	**5 out of 8**	**6 out of 8**	**Most Declared 3D Level:**
Binary Success:	**.87%**	**1.0%**	**.75%**	**.62%**	**.75%**	**Not Fam.**

Table 11: For each task, the participants had to provide the correct answer (see Table 9 for the individual tasks). The correct answers are colored in green while incorrect answers are colored in red. The binary success is calculated by giving correct answers the numerical value 1 and incorrect 0. Averaging turns it into percent. HW Dataset = Height and Weight dataset. PosCorr = positive correlation. NeuCorr = neutral correlation. NegCorr = negative correlation. Not Fam. = not familiar. Dece Fam. = decently familiar. Very Fam. = very familiar.

When examining each task, it can be noticed that the two most difficult tasks were Task 3 (68/100 difficulty score) and Task 4 (62/100 difficulty score). The two easiest tasks were Task 1 and 2

with respectively a task difficulty score of 37/100 and 25/100. Furthermore, the assumed task difficulty does not exactly pair up with how the participants rated the tasks. Only Task 1 and 4 received the expected difficulty level. Task 2 was easier than assumed, while Task 3 was found more difficult. Task 5 is a bit contradicting as out of the eight participants three found it easy to solve, but another three participants found it hard. As a result, the assumed task difficulty is not completely off, however, using the task difficulty score as guidance, the assumed task difficulty for future iterations should be stated as *Medium*.

	Red Wine Dataset	**Housing Dataset**	**Movie Dataset**	**Iris Dataset**	**HW Dataset**
Participant	**Task 1**	**Task 2**	**Task 3**	**Task 4**	**Task 5**
1	Medium	Easy	Hard	Hard	Hard
2	Hard	Medium	Hard	Medium	Easy
3	Medium	Easy	Hard	Medium	Hard
4	Medium	Hard	Medium	Medium	Medium
5	Easy	Medium	Hard	Medium	Easy
6	Medium	Easy	Easy	Medium	Medium
7	Easy	Easy	Medium	Medium	Easy
8	Easy	Easy	Medium	Hard	Hard
Assumed Task Difficulty:	**Medium**	**Medium**	**Medium**	**Medium**	**Hard**
Task Difficulty Score:	**37/100**	**25/100**	**68/100**	**62/100**	**50/100**
Most Declared Difficulty:	**Medium: 4 out of 8**	**Easy: 5 out of 8**	**Hard: 4 out of 8**	**Medium: 6 out of 8**	**Easy/Hard: 3 out of 8**

Table 12: Shows the the participants aggregated subjective rating of each tasks difficulty. The second to last row shows the difficulty level assumed by the researchers. The last row represents the calculated task difficulty score, which is calculated by averaging the participant ratings which are categorized as Easy (0), Medium (50), and Hard (100). HW Dataset = Height and Weight dataset.

5.3.3 Discussion

In the second iteration we saw a decline in our System Usability Scale score. We thought this was due to adding many new features that was not as refined and polished, as they could be. In this iteration we sought to improve this score by adding ideally better features and datasets, and by refining and polishing the existing ones. Moreover, as we have developed two versions of the application, we sought to know whether or not the new version functioned as intended. We did this by only conducting a usability test on our full AR integrated version (dynamic), and then internally tested the limited AR integrated version (static). Our thought was, if the dynamic version functioned as intended, then the static version would not show any implications.

Having examined the numbers in Section 5.3.2.2, it can be seen that the prototype did receive a better SUS score than the prior iteration. This is found to be an adequate result, especially due to the fact that two out of the eight participants greatly reduced the score. This together with them being the worst and second worst performing participant's goes well in-hand with them scoring the usability lower than the rest (see Table 11). Furthermore, as the most declared 3D level was

Not Familiar, and that the majority of participants found the prototype sufficient, it can be stated that the prototype is in a stage where it is final-test ready. However, before the final testing can start, we can readjust the assumed task difficulty as this iteration showed that some tasks were easier as well as others being more difficult than firstly anticipated (see Table 12). Although this does not directly influence the final testing, it is worthwhile to examine, as if the tasks were all found to be too difficult, then we would have to rethink and redesign the tasks to create a more balanced array of tasks.

Moreover, it is worth noting the Tasks 3, 4, 5 received the most incorrect answers. When compared to Tasks 1 and 2, Tasks 3, 4, and 5 all contain different aspects in terms of how to solve the task. Tasks 1 and 2 are both solved correctly by stating a positive correlation. Although this is also the case of Task 4, the task includes another axis. While Tasks 1 and 2 only required the participant to use two dimensions to solve the task, Task 4 involves a third dimension as well. As a result, the task becomes more complex to solve and this can be noticed in the results (see Table 11). Additionally, Task 3 presents a different way of solving a task. For this task, the participants had to locate the movie *The Dark Knight* in a large dataset (see Table 9). They did this by highlighting the data points one by one, and by using the timeline they could estimate where the specific data point was located. Due to the need of navigating differently compared to the previous tasks, and needing to be very specific, two participants were unsure in how to locate the data point. This could become an issue in the final test, however, with more proficient 3D visualization users the chances are limited. To answer Task 5, you must provide the correct estimation. This task was found to be equally easy as difficult. However, six out of eight provided the correct answer and those who got the incorrect answer, stated the task to be *hard*. Consequently, with a population with more proficiency in 3D visualizations, this task might lie between an *easy* to *medium* difficulty.

6 Experimental Design

Through the Analysis, User Research, and Design and Implementation (see Chapters 2, 4, and 5) we sought to understand the overarching complexities in data visualization, augmented reality, and cognitive load. To combat these complexities we will make use of reliable frameworks and try to form an experiment that helps us answer our two research questions. As outlined in Section 2.4, the two research questions aim to address two parts of designing and implementing an application. One part examines the development and user experience, and the other part focuses on measuring cognitive load - Both in context of an augmented reality data visualization application.

As for any other research project, it is important to consider the experimental variables. In this project, we have designed and developed two versions of our 3D data visualization application. The experimental version that includes a full augmented reality experience, and a control version that contains limited augmented reality capabilities. Consequently, the independent variable will be within the experimental group, as the users get exposed to the version with full AR integration. We will measure the difference between the two groups and their variables to figure out whether or not augmented reality helps to limit cognitive strain (specifically, measuring a difference in extraneous cognitive load) as well as enhance user experience in 3D data visualization.

In Chapter 3 we state that a total of 58 users participated in the experiment. In the early phases of the project we focused on the users ability and knowledge in statistics. As the experiment iterated over time, we changed from recording the participants statistical knowledge to recording their experience in 3D data visualization instead. The reason for this was grounded in statistical knowledge being so broad and containing so much information, a lot of which would be redundant in scope of this project, as opposed to 3D data visualization which is a more narrow and specific area which also mostly falls under the umbrella of statistical knowledge. The majority of the participants for this project stem from the Technical Faculty of Aalborg University in Copenhagen: a faculty with footprints in electronics, architecture, design, and planning, and were all sampled regardless of their level of 3D data visualization experience.

To answer the two research questions, data collection was needed. Our independent variable could not be compared with the dependent variable without testing two versions of the application. As a result, we chose a between-subjects design. Two groups who only got exposed to one version of the application. Both groups had a total of 29 users (experimental group (n = 29), control group (n = 29)). The experimental procedure consisted of 19 steps, with the testing session's duration varying between 20-25 minutes per participant. The experimental design protocol can be seen in 10.3 Appendix C. It includes a step-by-step walk-through of the experimental design used during the final testing session.

As explained throughout the report, we will utilize physiological data and to do so, we must gather it. Ethical considerations must be taken into account as physiological data can be sensitive.

Especially, BVP (or BPM when converted) can provide sensitive information about an individual. To mitigate this, we ask the participants to fill in an informed consent form. It includes information about the experience and their rights and data treatment. Moreover, a participant is able to withdraw their consent at any time. Each participant will be provided with an anonymized ID number to protect their private or sensitive information.

Planning to use reliable frameworks such as: UEQ and NASA-TLX for analyzing our data (see Sections 2.3.2.5 and 2.3.4.1), we aim for high internal validity of our results. Splitting our participants and the experiment into two groups allows us to compare the outcomes between the experimental group (dynamic) and control group (static). These two groups would of course be sampled from the same population to maximize optimal internal validity. Furthermore, having created a standardizing procedure (see 10.3 Appendix C) helps to ensure consistency in the collected data, and further helps us to control any extraneous variables. This together with precisely defining the independent and dependent variables and how to operationalize them reduces ambiguity and strengthens manipulations and measurements. As the final straw, pilot testing would assist in discovering any occurring issues in the experimental design. Conducting a pilot test on a small sample will be a priority before the final testing session.

In terms of external validity it could be argued that the project has got some limitations. By use of convenience sampling, the sample size might not have been accurately representative of the target group population, which decreased the experiments external validity. Despite using frameworks that can be used in a plethora of combinations (UEQ and NASA-TLX) our experimental design does not necessarily fit well into other experiments. If another experiment were to measure cognitive load and user experience, the same measures and procedures could be reused. However, this requires the experiment to follow our task-based experiment design. Without it, the NASA-TLX framework becomes redundant. Regardless of testing our prototype and the consistency of the frameworks used for our experiment, we have not replicated the testing scenario in a different location, nor have the experimental design been used in a different setting, with different populations, or methodologies. Consequently, field experiments have not been conducted outside the comfort of our testing location, and the experimental design has only been tested on one population. As a result, the external validity of this experimental design becomes very limited, and we cannot guarantee that the research study will apply to other settings.

Based on our experimental design, we expect our findings to show promising signs in the user experience and cognitive load department. We anticipate that the full augmented reality integrated version of our application (experimental group), receives a better user experience and cognitive load outcomes than the static and limited version (control group).

7 Findings

The findings of our experimental design will be presented in this section. This involves: User Experience, Task Performance, and Physiological data. Answering the two research questions (see Section 2.4) will be based on these findings. Only a presentation of the data will be presented in this section. Find the analysis of the data in Section 8.

7.1 User Experience

For measuring user experience we made use of the User Experience Questionnaire (UEQ). This framework provided us with data on user experience divided into six sub-scales (see Section 2.3.2.5 for more information on the sub-scales). We have calculated a comparison of scale means for both the experimental group (EG) and control group (CG) users (see Tables 13 and 14). Although testing our application on the desired target group, we have a generally low Confidence percentage that the results represent the true population. As a result, the Confidence Interval's can be misleading or at least not represent the true population as the Confidence levels vary between $\approx 30 - 40\%$. This leaves too big of a room to fill as we used a 5% confidence interval for both groups. That being stated, the Mean scores for both groups show a relatively positive outcome. As the UEQ utilize a 7-point Likert Scale, the values were transformed into positive and negative values varying from -3 to +3. As a result, the maximum score a sub-scale could achieve was a Mean of 3.00. With all Mean scores being above 1.00, despite of the Novelty sub-scale in the control group (0.77), it can be stated that the user experience in both versions has been satisfactory.

More interesting is to compare the two versions in a Two Sample T-Test. As could be derived from the Mean scores, it was difficult to differentiate between the two groups. However, noticeable is the difference in the Stimulation and Novelty categories between the two groups. The experimental group scored significantly higher in both the Stimulation and Novelty categories. This can be seen in Table 15 with the Stimulation sub-scale obtaining a P-Value of 0.0348 and the Novelty sub-scale obtaining a P-Value of 0.0195. With the Alpha-Level being set to 0.05, this shows a significant difference between the two versions of the application in the Stimulation and Novelty department. The reason for this will be further analyzed in the Discussion of the report (see Section 8).

Experimental Group (n = 29) Comparison of Scale Means (UEQ)					
Scale	**Mean**	**STD**	**Confidence**	**Confidence Interval**	
Attractiveness	1.43	0.82	0.30	1.13	1.72
Perspicuity	1.10	0.94	0.34	0.76	1.45
Efficiency	1.35	0.93	0.34	1.01	1.69
Dependability	1.26	0.72	0.26	1.00	1.52
Stimulation	1.53	0.66	0.24	1.29	1.77
Novelty	1.33	0.80	0.29	1.04	1.62

Table 13: Shows the scale means and the corresponding 5% confidence intervals for the dynamic group with full AR integration (experimental group).

Control Group (n = 29) Comparison of Scale Means (UEQ)					
Scale	**Mean**	**STD**	**Confidence**	**Confidence Interval**	
Attractiveness	1.01	1.09	0.40	0.61	1.40
Perspicuity	1.33	1.15	0.42	0.91	1.74
Efficiency	1.35	0.99	0.36	0.99	1.71
Dependability	1.37	0.95	0.34	1.03	1.72
Stimulation	1.05	0.97	0.35	0.70	1.41
Novelty	0.77	0.97	0.35	0.42	1.12

Table 14: Shows the scale means and the corresponding 5% confidence intervals for the static group with limited AR integration (control group).

Experimental group vs. Control Group Two Sample T-Test (UEQ)		
Scale	**P-Value**	**Alpha-Level: 0.05**
Attractiveness	0.1034	No Significant Difference
Perspicuity	0.4193	No Significant Difference
Efficiency	1.0000	No Significant Difference
Dependability	0.6141	No Significant Difference
Stimulation	0.0348	Significant Difference
Novelty	0.0195	Significant Difference

Table 15: Shows a two sample t-test to check if the scale means of the two measured versions of the application differ significantly. The Alpha-Level is 0.05.

7.2 Task Performance

In Section 2.3.4.4 we mention *Task on Time* and *Task Success Rate* could act as a measure for cognitive load. We assumed having spent more time on a task equaled a more mentally demanding task. Furthermore, we assumed that answering incorrectly would imply higher levels of cognitive load required for completing a task successfully. Having collected and measured this data in the final testing session, we have gained an insight into whether this assumption is true or not. Tables 16 and 17 present the experimental (dynamic) and control (static) groups correct and incorrect answers from each of the five tasks. Worth noting is the binary success percentage in both Tables. The experimental group (EG) scored as follows: Task 1: 89%, Task 2: 93%, Task 3: 93%, Task 4: 93%, and Task 5: 93%. The control group's (CG) score was: Task 1: 89%, Task 2: 96%, Task 3: 100%, Task 4: 82%, and Task 5: 79%. This shows that the two groups are almost identical in their answers, although a small dip in performance is noticed from the CG users in Tasks 4 and 5. Additionally, we measured the users movement. This measure was based on the researchers' own perceptions of the users' movement during the tasks. Noticeable is that only the EG users' have got an average movement score. This is due to the CG users using a static setup and version of the application.

Experimental Group (n = 29) Dynamic - Full AR integration						
	Red Wine Dataset	**Housing Dataset**	**Movie Dataset**	**Iris Dataset**	**HW Dataset**	
Participant	**Task 1**	**Task 2**	**Task 3**	**Task 4**	**Task 5**	**3D Level**
1	PosCorr	PosCorr	Correct	PosCorr	4	Dece Fam.
3	PosCorr	NeuCorr	Correct	PosCorr	4	Not Fam.
4	PosCorr	PosCorr	Correct	Unsure	4	Dece Fam.
5	PosCorr	PosCorr	Correct	PosCorr	4	Not Fam.
6	PosCorr	PosCorr	Correct	PosCorr	4	Not Fam.
7	PosCorr	NegCorr	Correct	PosCorr	4	Dece Fam.
21	PosCorr	PosCorr	Correct	PosCorr	4	Dece Fam.
22	PosCorr	PosCorr	Correct	PosCorr	4	Dece Fam.
23	PosCorr	PosCorr	Correct	PosCorr	4	Dece Fam.
24	PosCorr	PosCorr	Unsure	PosCorr	4	Dece Fam.
25	PosCorr	PosCorr	Correct	PosCorr	4	Dece Fam.
26	PosCorr	PosCorr	Correct	PosCorr	4	Not Fam.
27	PosCorr	PosCorr	Correct	PosCorr	3	Not Fam.
28	PosCorr	PosCorr	Correct	PosCorr	4	Not Fam.
29	PosCorr	PosCorr	Correct	PosCorr	4	Dece Fam.
30	PosCorr	PosCorr	Correct	PosCorr	4	Not Fam.
47	NegCorr	PosCorr	Correct	MixedCorr	4	Not Fam.
48	PosCorr	PosCorr	Correct	PosCorr	4	Dece Fam.
49	PosCorr	PosCorr	Correct	PosCorr	4	Not Fam.
50	PosCorr	PosCorr	Correct	PosCorr	4	Not Fam.
51	PosCorr	PosCorr	Correct	PosCorr	4	Very Fam.
52	MixedCorr	PosCorr	Correct	PosCorr	4	Dece Fam.
53	PosCorr	PosCorr	Correct	PosCorr	4	Dece Fam.
54	PosCorr	PosCorr	Correct	PosCorr	4	Very Fam.
55	PosCorr	PosCorr	Correct	PosCorr	4	Not Fam.
56	PosCorr	PosCorr	Correct	PosCorr	4	Dece Fam.
57	NegCorr	PosCorr	Correct	PosCorr	4	Dece Fam.
58	PosCorr	PosCorr	Correct	PosCorr	3	Dece Fam.
59	PosCorr	PosCorr	Unsure	PosCorr	4	Not Fam.
Correct Answers:	**26 out of 29**	**27 out of 29**	**27 out of 29**	**27 out of 29**	**27 out of 29**	**Most Declared 3D Level:**
Binary Success:	**.89%**	**.93%**	**.93%**	**.93%**	**.93%**	**Dece Fam.**
Avg. Movement:	**3.8**	**3.2**	**4.0**	**3.8**	**4.3**	**15 out of 29**

Table 16: For each task, the participants had to provide the correct answer (see Table 9 for the individual tasks). The correct answers are colored in green while incorrect answers are colored in red. The binary success is calculated by giving correct answers the numerical value 1 and incorrect 0. Averaging turns it into percent. Avg. Movement is based on a numerical value of 0-5, where 0 is no movement at all and 5 is substantial movement, e.g. moving around and crouching to change perspective during the completion of the tasks. HW Dataset = Height and Weight dataset. PosCorr = positive correlation. NeuCorr = neutral correlation. NegCorr = negative correlation. MixedCorr = mixed correlation. Not Fam. = not familiar. Dece Fam. = decently familiar. Very Fam. = very familiar.

Static Group (n = 29) Static - Limited AR Integration						
	Red Wine Dataset	Housing Dataset	Movie Dataset	Iris Dataset	HW Dataset	
Participant	Task 1	Task 2	Task 3	Task 4	Task 5	3D Level
8	PosCorr	PosCorr	Correct	PosCorr	2	Not Fam.
9	PosCorr	PosCorr	Correct	PosCorr	4	Not Fam.
10	PosCorr	PosCorr	Correct	MixedCorr	2	Not Fam.
11	PosCorr	NeuCorr	Correct	PosCorr	4	Not Fam.
12	PosCorr	PosCorr	Correct	PosCorr	3	Dece Fam.
13	NeuCorr	PosCorr	Correct	PosCorr	4	Not Fam.
14	PosCorr	PosCorr	Correct	PosCorr	4	Dece Fam.
15	PosCorr	PosCorr	Correct	PosCorr	4	Dece Fam.
16	PosCorr	PosCorr	Correct	PosCorr	4	Dece Fam.
17	PosCorr	PosCorr	Correct	Unsure	4	Not Fam.
18	PosCorr	PosCorr	Correct	PosCorr	4	Not Fam.
19	PosCorr	PosCorr	Correct	PosCorr	3	Dece Fam.
20	PosCorr	PosCorr	Correct	PosCorr	4	Dece Fam.
31	NegCorr	PosCorr	Correct	PosCorr	4	Not Fam.
32	PosCorr	PosCorr	Correct	PosCorr	4	Not Fam.
33	PosCorr	PosCorr	Correct	PosCorr	4	Not Fam.
34	PosCorr	PosCorr	Correct	MixedCorr	4	Dece Fam.
35	PosCorr	PosCorr	Correct	PosCorr	4	Very Fam.
36	PosCorr	PosCorr	Correct	MixedCorr	4	Very Fam.
37	PosCorr	PosCorr	Correct	PosCorr	4	Dece Fam.
38	PosCorr	PosCorr	Correct	PosCorr	4	Very Fam.
39	NegCorr	PosCorr	Correct	MixedCorr	3	Dece Fam.
40	PosCorr	PosCorr	Correct	PosCorr	3	Dece Fam.
41	PosCorr	PosCorr	Correct	PosCorr	4	Very Fam.
42	PosCorr	PosCorr	Correct	PosCorr	4	Not Fam.
43	PosCorr	PosCorr	Correct	PosCorr	4	Dece Fam.
44	PosCorr	PosCorr	Correct	PosCorr	4	Not Fam.
45	PosCorr	PosCorr	Correct	PosCorr	3	Dece Fam.
46	PosCorr	PosCorr	Correct	PosCorr	4	Dece Fam.
Correct Answers:	**26 out of 29**	**28 out of 29**	**29 out of 29**	**24 out of 29**	**23 out of 29**	**Most Declared 3D Level:**
Binary Success:	**.89%**	**.96%**	**1.0%**	**.82%**	**.79%**	**Dece Fam.**
Avg. Movement:	**0**	**0**	**0**	**0**	**0**	**13 out of 29**

Table 17: For each task, the participants had to provide the correct answer (see Table 9 for the individual tasks). The correct answers are colored in green while incorrect answers are colored in red. The binary success is calculated by giving correct answers the numerical value 1 and incorrect 0. Averaging turns it into percent. Avg. Movement is based on a numerical value of 0-5, where 0 is no movement at all and 5 is substantial movement, e.g. moving around and crouching to change perspective during the completion of the tasks. HW Dataset = Height and Weight dataset. PosCorr = positive correlation. NeuCorr = neutral correlation. NegCorr = negative correlation. MixedCorr = mixed correlation. Not Fam. = not familiar. Dece Fam. = decently familiar. Very Fam. = very familiar.

Additionally, data on the users' perception on task difficulty was obtained. The assumed task difficulty was based off of the prior iteration's result (see Table 12). The users' perception of the

five tasks difficulty in the experimental group was as follows: Task 1: 30/100, Task 2: 13/100, Task 3: 39/100, Task 4: 44/100, and Task 5: 29/100. The control group scored the tasks: Task 1: 20/100, Task 2: 10/100, Task 3: 31/100, Task 4: 60/100, and Task 5: 51/100. The three most noticeable task difficulty scores are the differences in Tasks 1, 4, and 5. The EG users scored Task 1 14 points higher than the CG users in Task 1. However, the EG users scored the difficulty 16 and 22 points lower in Tasks 4 and 5 than the CG users.

Experimental Group: (n = 29) Dynamic - Full AR Integration					
	Red Wine Dataset	**Housing Dataset**	**Movie Dataset**	**Iris Dataset**	**HW Dataset**
	Task 1	**Task 2**	**Task 3**	**Task 4**	**Task 5**
Assumed Task Difficulty:	Medium	Easy	Hard	Medium	Medium
Task Difficulty Score:	**34/100**	**13/100**	**39/100**	**44/100**	**29/100**
Most Declared Difficulty:	**Medium: 14 out of 29**	**Easy: 23 out of 29**	**Medium: 19 out of 29**	**Medium: 18 out of 29**	**Easy: 15 out of 29**

Table 18: Shows the the participants aggregated subjective rating of each tasks difficulty in the experimental group. The second to last row shows the difficulty level assumed by the researchers. The last row represents the calculated task difficulty score, which is calculated by averaging the participant ratings which are categorized as Easy (0), Medium (50), and Hard (100). HW Dataset = Height and Weight dataset.

Control Group: (n = 29) Static - Limited AR Integration					
	Red Wine Dataset	**Housing Dataset**	**Movie Dataset**	**Iris Dataset**	**HW Dataset**
	Task 1	**Task 2**	**Task 3**	**Task 4**	**Task 5**
Assumed Task Difficulty:	Medium	Easy	Hard	Medium	Medium
Task Difficulty Score:	**20/100**	**10/100**	**31/100**	**60/100**	**51/100**
Most Declared Difficulty:	**Easy: 19 out of 29**	**Easy: 24 out of 29**	**Medium: 14 out of 29**	**Medium: 15 out of 29**	**Medium: 19 out of 29**

Table 19: Shows the the participants aggregated subjective rating of each tasks difficulty in the control group. The second to last row shows the difficulty level assumed by the researchers. The last row represents the calculated task difficulty score, which is calculated by averaging the participant ratings which are categorized as Easy (0), Medium (50), and Hard (100). HW Dataset = Height and Weight dataset.

As a method for measuring cognitive load we utilized the NASA-TLX framework. We calculated the NASA-TLX score for each individual task as well as an aggregated overall NASA-TLX score. Furthermore, we measured the users' task duration for each task. Tables 20 and 21 present this data. Generally, the experimental group spent more time than the control group on each individual task. Likewise, the EG users showed a greater Task Load Index for each individual task when compared to the CG users. By looking at the numbers alone, this confirms our assumption that spending more time on a task demands more mental workload. However, nothing can with 100% certainty be confirmed on these numbers alone. A more in-depth analysis is required to significantly accept this assumption.

Experimental Group (n = 29) Task Duration and Aggregated NASA-TLX Score					
	Task 1	**Task 2**	**Task 3**	**Task 4**	**Task 5**
Task Duration:	109.5s	74.4s	91.7s	129.9s	121.3s
Time Difference to CG:	+12.9s	+12.9s	+23.0s	+18.1s	+37.6s
NASA-TLX Score:	44.13	38.76	44.13	49.56	44.67
Aggregated Score:	**44.29 out of 100**				

Table 20: The task duration and NASA-TLX Scores for each task and aggregated score for the dynamic group with full AR integration (experimental group).

Control Group (n = 29) Task Duration and Aggregated NASA-TLX Score					
	Task 1	**Task 2**	**Task 3**	**Task 4**	**Task 5**
Task Duration:	96.6s	61.5s	68.7	111.8s	83.7s
Time Difference to EG:	-12.9s	-12.9s	-23.0s	-18.1s	-37.6s
NASA-TLX Score:	40.34	34.23	42.66	50.19	40.49
Aggregated Score:	**41.58 out of 100**				

Table 21: The task duration and NASA-TLX Scores for each task and aggregated score for the static group with limited AR integration (control group).

In Table 22, the significance test results for each task are shown, including the test statistic, p-value, and the test used. The data was checked for normality, and a paired t-test was used if the data was normally distributed, while the Mann-Whitney U test was used if the data was not normally distributed. The NASA-TLX scores are weighted and aggregated for each participant, and significance tests are conducted to determine differences between the experimental and control groups. Overall, the results suggest that there were no statistically significant differences in the NASA-TLX scores between the experimental and control groups across all tasks. This indicates that the perceived workload and cognitive load experienced by participants, according to the NASA-TLX, were comparable between the groups for each task, despite what AR application they used.

NASA-TLX Significance Test Results			
Task	**Statistic**	**P-value**	**Test Used**
Task 1	1.206	0.233	T-Test
Task 2	1.512	0.203	Mann-Whitney U
Task 3	0.408	0.685	T-Test
Task 4	-0.116	0.908	T-Test
Task 5	1.207	0.232	T-Test

Table 22: Significance test results for each task (experimental vs control group), including the test statistic, p-value, and the test used. The data was checked for normality, and a paired t-test was used if the data was normally distributed, while the Mann-Whitney U test was used if the data was not normally distributed.

7.3 Physiological Data

In this section, the physiological data collected during the final test, including Blood Volume Pulse (BVP) and Electrodermal Activity (EDA) will be presented and described.

7.3.1 Blood Volume Pulse (BVP)

Table 23 displays the mean values of the BVP measures, including Beats Per Minute (BPM), Standard Deviation of Normal-to-Normal Intervals (SDNN), and Root Mean Square of Successive Differences (RMSSD) for the baseline and each task, and for both the experimental and control group. As noted in the Methodology Chapter (see Chapter 3), the baseline values have been subtracted from the Beats Per Minute (BPM), which are shown as the percentage increase relative to the baseline, with higher values indicating an increase in physical or mental stress levels. The experimental group (dynamic) experienced the highest increase in BPM relative to the baseline in all tasks, indicating that the dynamic version had a greater impact on their heart rate. Notably, the experimental group exhibited an increase of over 7% BPM from baseline to Task 2, which was the most stressful task, while the control group's BPM is relatively small, suggesting that the dynamic version clearly has a stronger impact on the experimental groups heart rate. Tasks 1 and 4 also revealed a considerable difference in BPM between the two groups. The Mann-Whitney U test results for BPM show a trend towards significance in Task 2 ($p = 0.0538$), suggesting that the differences in BPM between the groups may be meaningful, especially as the experimental group consistently shows higher BPM values (see Table 23). SDNN measures heart rate variability, with higher values indicating enhanced cardiovascular fitness and autonomic flexibility. The experimental group demonstrated higher SDNN values than the control group, implying superior autonomic control during tasks. The experimental group had higher SDNN values in all tasks, although some values were quite close. Specifically, in Task 1 and Task 5, the experimental group showed a remarkably higher SDNN compared to the control group, reflecting a more relaxed state overall (see Table 23). Notably, The Mann-Whitney U test results for SDNN show significant differences in Task 1 ($p = 0.0717$), Task 3 ($p = 0.0046$) and Task 5 ($p = 0.0620$), indicating that the experimental group

had significantly better heart rate variability compared to the control group during these tasks. Similarly, RMSSD also showed higher values for the experimental group compared to the control group in several tasks indicating better parasympathetic activity and stress resilience during tasks. However, in Task 2, the experimental group achieved worse heart rate variability than the control group. Interestingly, the Mann-Whitney U test results showed there were no significant differences between the two groups for RMSSD which suggest that while the experimental group consistently maintained higher RMSSD values, the differences were not statistically significant across all tasks.

Overall, the experimental group consistently showed higher BPM and better SDNN, and RMSSD values across tasks, suggesting they experienced greater physiological arousal and maintained better autonomic regulation using the dynamic version of the application.

Experimental (n = 29) and Control Group (n = 29) Blood Volume Pulse (BVP)						
	BPM Mean (Baseline Substracted)		**SDNN - Mean**		**RMSSD - Mean**	
	Experimental Group	**Control Group**	**Experimental Group**	**Control Group**	**Experimental Group**	**Control Group**
Baseline:	**-**	**-**	**63.06**	**52.99**	**39.91**	**34.38**
Task 1	24.19%	20.35%	63.06	52.99	39.91	34.38
	Mann-Whitney: 0.1570		**Paired T-Test: 0.0717**		**Mann-Whitney: 0.1314**	
Task 2	27.36%	19.95%	61.35	61.18	42.18	46.47
	Mann-Whitney: 0.0538		**Mann-Whitney: 0.9133**		**Mann-Whitney: 0.8641**	
Task 3	23.49%	19.07%	61.21	54.94	36.92	37.66
	Paired T-Test: 0.1462		**Paired T-Test: 0.2691**		**Mann-Whitney: 1**	
Task 4	27.93%	22.01%	61.77	57.04	40.18	38.56
	Mann-Whitney: 0.1024		**Mann-Whitney: 0.9256**		**Mann-Whitney: 0.4368**	
Task 5	25.00%	22.60%	58.72	51.41	37.43	36.15
	Paired T-Test: 0.4258		**Mann-Whitney: 0.0620**		**Mann-Whitney: 0.3120**	

Table 23: The following table displays the mean values of Blood Volume Pulse (BVP) measures: Beats Per Minute (BPM), Standard Deviation of Normal-to-Normal Intervals (SDNN), and Root Mean Square of Successive Differences (RMSSD) for the baseline and each task in the experimental group with full Augmented Reality (AR) integration. BPM is represented as the percentage increase relative to the baseline with the baseline subtracted. The significance of the difference between the experimental and control groups was tested for normal distribution, using a paired t-test if normally distributed, or the Mann-Whitney U test if not, with a significance threshold of $p < 0.05$

Figure 40 visualizes the physiological data related to Beats Per Minute (BPM), Standard Deviation of Normal-to-Normal Intervals (SDNN), and Root Mean Square of Successive Differences

(RMSSD) across the different tasks and baseline for both the experimental and control groups. The figure displays three plots for each of the three BVP measures used and shows the mean of each measure across the baseline and the different tasks. This essentially shows the same data as Table 23, just visualized. The first plot presents the BPM values, shown as the percentage increase relative to the baseline. The second plot illustrates SDNN values, representing heart rate variability. Both groups show a decrease in SDNN from baseline to Task 1, with the experimental group maintaining higher values throughout the tasks. However, the control group experiences a sharper decline, especially noticeable in Task 2 and Task 4, suggesting increased stress and reduced heart rate variability for these tasks. The third plot represents RMSSD values, focusing on short-term heart rate variability. The experimental group exhibits higher RMSSD values compared to the control group in three of the five tasks. However, the control group had a significantly better RMSSD value than the experimental group in Task 2. Both SDNN (see Figure 40, Middle) and RMSSD (see Figure 40, Right) exhibit very low heart rate variability during Task 1 and Task 5 when compared to the other tasks.

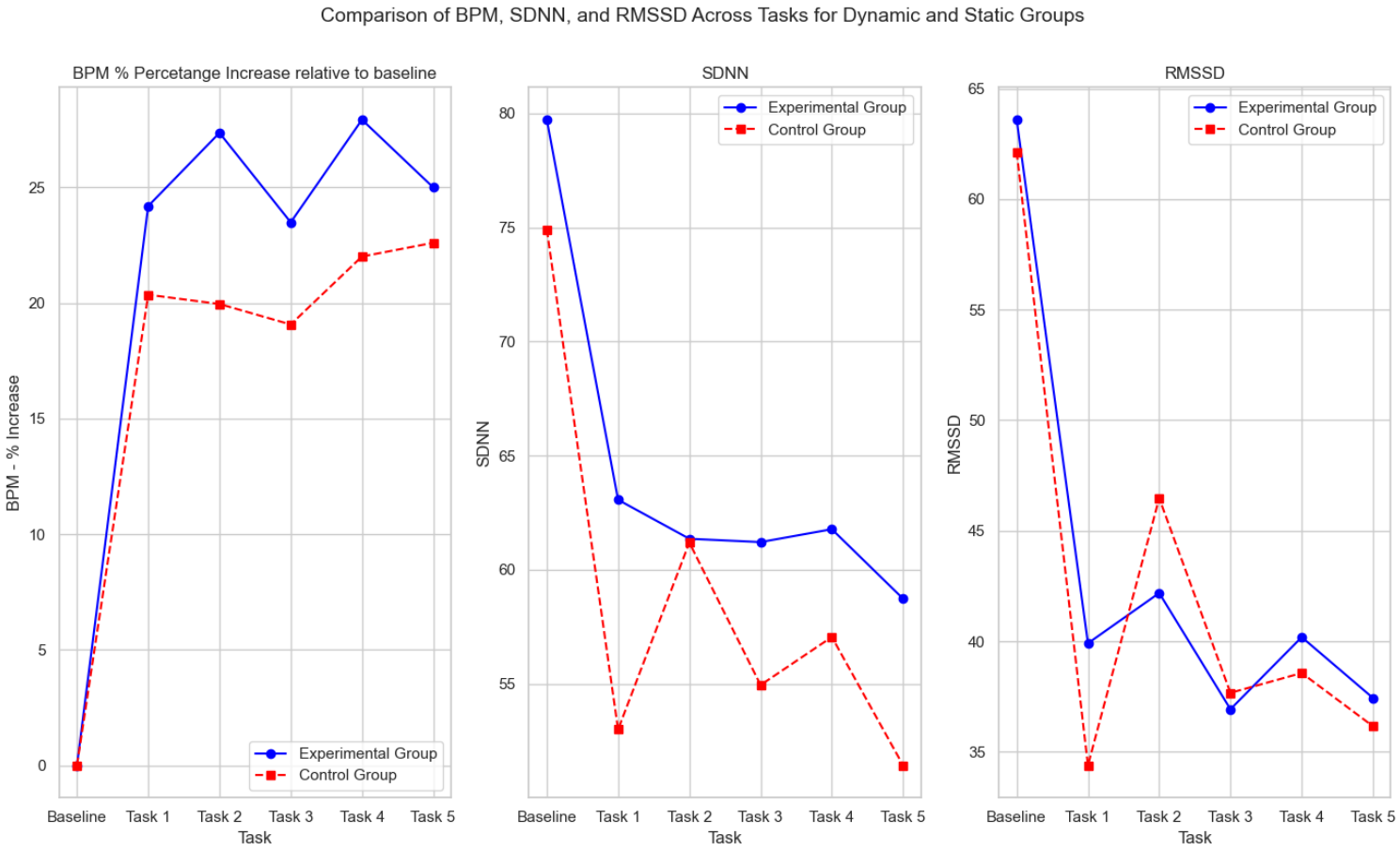


Figure 40: Comparison of BPM, SDNN, and RMSSD across tasks for Experimental and Control Groups. The figure displays three plots for each of the three BVP measures used, showing the mean values across the baseline and the different tasks. BPM is represented as the percentage increase relative to the baseline with the baseline subtracted. For SDNN and RMSSD, lower values indicate worse heart rate variability and higher levels of stress or strain in participants.

Overall, the experimental group consistently shows better BPM, SDNN, and RMSSD values across tasks, suggesting they experienced greater physiological arousal and maintained better autonomic regulation during the dynamic version of the application. Task 4 is identified as the most stressful, with significant increases in BPM and decreases in SDNN and RMSSD, indicating increased stress and reduced heart rate variability. Task 2 also stands out with substantial changes in SDNN and RMSSD.

7.3.2 Electrodermal Activity (EDA)

As previously mentioned in the Methodology Chapter (see Chapter 3), Electrodermal Activity (EDA) measures skin conductance changes and provides insights into emotional arousal and cognitive load. Table 24 displays the mean values of EDA measures which consists of baseline-subtracted EDA mean, The number of skin Conductance Response (SCR) Peaks, and SCR Peaks

Amplitude for the baseline and each of the five tasks and for both experimental and control group. The raw EDA mean indicates the overall change in skin conductance, shown as the percentage increase relative to the baseline, with the baseline subtracted. The experimental group had higher EDA means compared to the control group, with notable increases during all tasks (see Table 24), except in Task 5. Moreover, the experimental group achieved significantly higher raw EDA mean values, especially in Task 1, where the Mann-Whitney U Test almost proved significant ($p = 0.0664$). Although the experimental group achieved a higher percentage increase in BPM relative to the baseline across all tasks, Tasks 2-5 are not statistically significant. The number of SCR peaks during a task reflects the frequency of skin conductance responses. It should be noted that this number of peaks has been divided and normalized by the duration of the tasks. The experimental group showed a significantly higher SCR peak count across all tasks, with the experimental group achieving almost double the peaks in Task 2 ($p = 0.0159$) and Task 3 ($p = 0.004$). In Tasks 1, 3, 4, and 5, the experimental group experienced about 50%-75% more peaks compared to the control group. In general, the experimental groups achieved a 150% rise in number of SCR peaks during the tasks in contrast to the baseline, while the control group achieved 60-65%. The Mann-Whitney U Test results for SCR Peak Count indicate significant differences in Task 2 ($p = 0.0159$), Task 3 ($p = 0.0046$), Task 4 ($p = 0.0002$), and Task 5 ($p = 0.0152$), highlighting the substantial difference in emotional arousal between the experimental and control groups. The SCR Peaks Amplitude provides insight into the intensity of the emotional responses, presented as the percentage increase relative to the baseline (see Table 24). The experimental group exhibited higher SCR Peaks Amplitude values compared to the control group, particularly in Task 1 and Task 3. The Mann-Whitney U Test results for SCR Peaks Amplitude in Task 1 ($p = 0.0064$) and Task 3 ($p = 0.0400$) also indicated significant differences, with the experimental group showing a lot higher amplitude values. However, there were no significant differences in Tasks 2, 4, and 5, suggesting that while the experimental group consistently maintained higher amplitude values, these differences were not statistically significant in those tasks.

Experimental (n = 29) and Control Group (n = 29) Electrodermal Activity (EDA)						
	EDA Mean		**SCR Peaks Count**		**SCR Peaks Amplitude**	
	Experimental Group	**Control Group**	**Experimental Group**	**Control Group**	**Experimental Group**	**Control Group**
Baseline:	**-**	**-**	**9.65**	**9.41**	**-**	**-**
Task 1	51.79%	26.95%	24.77	16.50	49.20%	6.80%
	Mann-Whitney: 0.0664		**Mann-Whitney: 0.0688**		**Mann-Whitney: 0.0064**	
Task 2	60.25%	38.39%	29.62	16.78	35.04%	23.28%
	Mann-Whitney: 0.1437		**Mann-Whitney: 0.0159**		**Mann-Whitney: 0.2192**	
Task 3	58.35%	44.12%	26.71	16.10	41.15%	10.03%
	Mann-Whitney: 0.3272		**Mann-Whitney: 0.0046**		**Mann-Whitney: 0.0400**	
Task 4	60.25%	44.36%	23.29	14.31	33.86%	31.23%
	Paired T-Test: 0.1805		**Mann-Whitney: 0.0002**		**Mann-Whitney: 0.7322**	
Task 5	51.96%	44.52%	24.82	16.11	20.59%	-3.30%
	Paired T-Test: 0.5790		**Mann-Whitney: 0.0152**		**Mann-Whitney: 0.1760**	

Table 24: The following table displays the mean values of Blood Volume Pulse (BVP) measures Beats Per Minute (BPM), Standard Deviation of Normal-to-Normal Intervals (SDNN), and Root Mean Square of Successive Differences (RMSSD) for the baseline and each task in the experimental group with full Augmented Reality (AR) integration. The number of skin conductance responses (SCR Peaks Count) have been normalized with the time each participant spent on a task.

The first plot (see Figure 41, Left) presents the average number of SCR Peaks across both the baseline and tasks, which indicates the frequency or amount of skin conductance responses per task/baseline. The dynamic group achieved a consistently higher number of SCR peaks compared to the static group across all tasks, with a very significant difference in amount of SCR peaks achieved on average by the participants in Task 2, which also was statistically significant according to the Mann-Whitney test (see Table 24) which could indicate greater emotional arousal and cognitive load. The second plot (see Figure 41, Middle) illustrates the SCR Peaks Amplitude Mean, reflecting the mean amplitude/intensity of all the SCR peaks that occurred during a task/session. It is represented as the percentage increase relative to the baseline, with the baseline subtracted. The experimental group exhibits higher amplitude mean values across most tasks, with significant peaks in Tasks 1 and 3. This also corresponds to the significant differences in SCR Peaks Amplitude observed in Tasks 1 and 3, with the experimental group showing more intense emotional responses compared to the control group. The third plot (see Figure 41, Right), portrays the Raw EDA Mean, indicating the overall change in skin conductance, with the percentage increase relative to the baseline, with the baseline subtracted. The dynamic group had slightly higher EDA means compared to the static group. Furthermore, the dynamic group had consistently higher

EDA means compared to the static group across all tasks, except in Task 5. This general trend reflects the greater emotional arousal and cognitive load experienced by the experimental group, although the values were not significant (see Table 24).

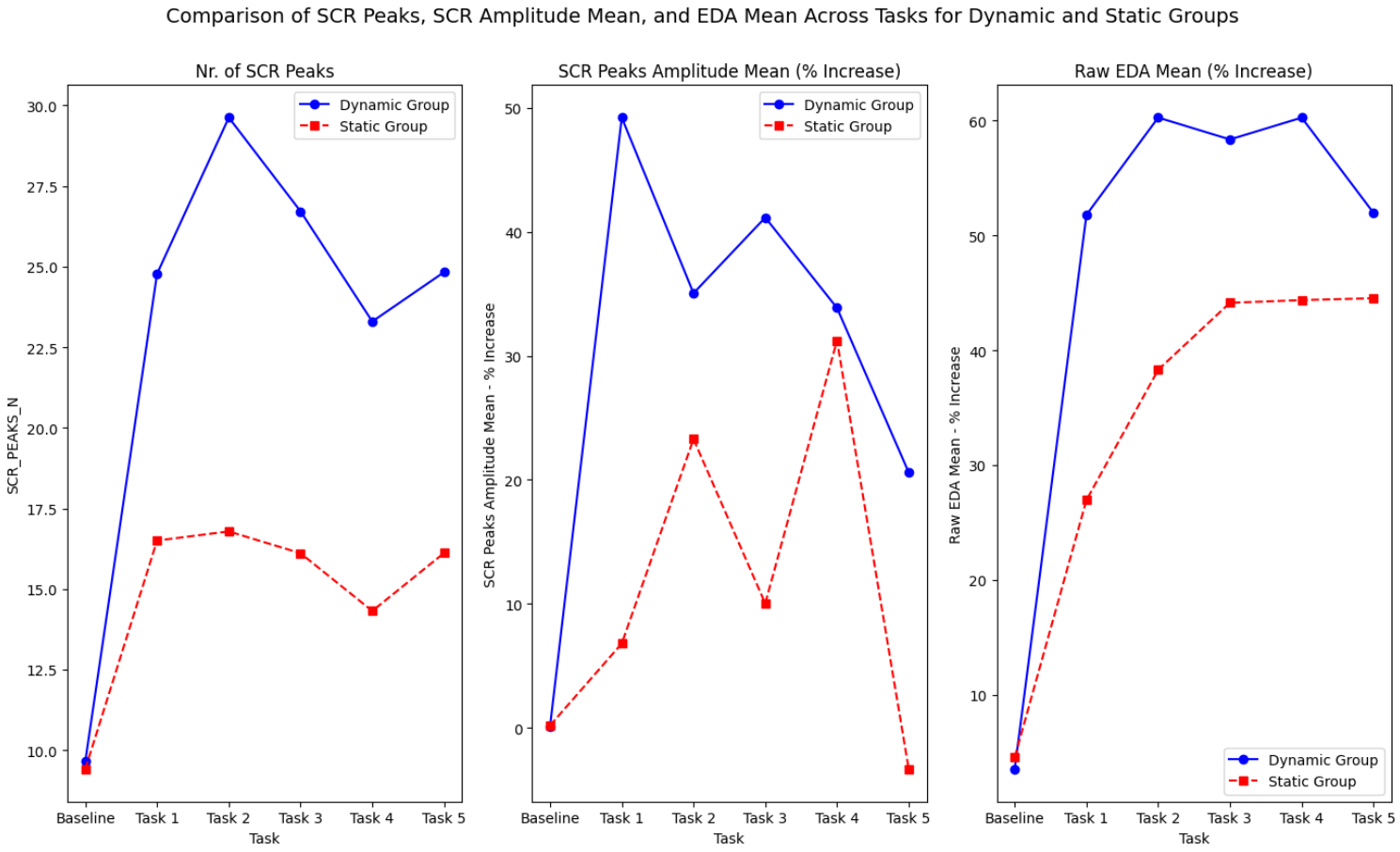


Figure 41: Comparison of SCR Peaks, SCR Amplitude Mean, and EDA Mean across tasks for Dynamic and Static Groups. The figure displays three plots showing the mean values across the baseline and different tasks for each measure. It is worth noting that the raw EDA mean and SCR Amplitude Mean are represented as the percentage increase relative to the baseline, with the baseline subtracted. The significance of the difference between the experimental and control groups was tested for normal distribution, using a paired t-test if normally distributed, or the Mann-Whitney U test if not, with a significance threshold of $p < 0.05$

8 Discussion

Throughout this project different approaches have been used (see Chapters 2 and 3). We decided that an iterative design would support us the best when it came to designing and implementing the application. More specifically, we chose to utilize the four steps of User-Centered Design, which are: *Understand context of use*, *Specify user requirements*, *Design solutions*, and *Evaluate against requirements* (see Section 2.3.2). Following this approach ensured a clearly defined context of use for the application as well as specific user requirements, which we in turn could evaluate upon. Consequently, we looked back on how we used the framework and if we could have done more to structurally follow it.

Conducting user research helped us to scope in on the important factors of statistics and data visualization. However, as we only interviewed a total of five users, we only managed to gather an exploratory view on the subject, and not a tendency in the users data approach patterns. While these interviews did provide us with knowledge that were used to form success criteria for the application (see Section 4.1.3), the information gathered from the qualitative work and further examinations of the users' reached a standstill. Although frequently establishing and conducting usability tests, we did not iteratively gather qualitative data, but only quantitative. If this had been done, more in-depth feedback would have been provided, which could have resulted in a more polished and refined application that fitted the users' exact needs. As the whole purpose of using User-Centered Design is to have the user in mind at any time, we might have neglected an in-depth analysis at some stages of the project, i.e. asking them how they visualize data, what their process is, etc. However, as we developed the prototype iteratively, which was followed up by testing sessions and evaluations, we did collect useful quantitative data often.

As the quantitative data was gathered through the System Usability Scale (SUS) questionnaire (see Sections 2.3.2.6 and 3.2), we primarily relied on evaluating our application's usability based on the SUS scale scores. Although this did provide comparable data from one iteration to another, it did not strengthen the already lacking qualitative department of the iterations. As Nielsen stated (see Section 2.3.2.6) usability tests should primarily focus on identifying usability problems, collecting qualitative data, and assessing participants' overall satisfaction with a product. While the System Usability Scale for the most part does check the boxes, gaining more in-depth information about a user's experience lacked a whole lot. Having conducted interviews more frequently with the users would for one: align with the User-Centered Design approach, and secondly provide qualitative data in which could result in a deeper user and product understanding. However, when it is all said and done, examining how well we managed to resolve the success criteria (see Section 4.1.3), is what should be in focus. Derived from the Interview Analysis (see Section 4.1.2) it was found that the users' wanted an application that was firstly; easy to use, secondly; with data visualizations that were easy to interpret, and thirdly; the learnability should not require much effort.

Having used an iterative approach resulted in us performing usability tests on multiple occasions. Examining the differences in SUS score across the three versions (see Tables 3, 6, and 10), showed that the application seemed to be in a state that can be easily interacted with. Additionally, it can be stated that the data visualization aspect was easy to interpret as users with non or lesser familiarity with 3D data visualization, managed to perform well in the task scenarios (see Tables 7, 11 16, and 17). This also goes well in-hand with the learnability as those with lesser familiarity in 3D data visualization and those with lesser statistical knowledge managed to solve the tasks without too many errors. The learnability could also be correlated to the aggregated NASA-TLX scores (see Tables 20 and 21). However, a more in-depth analysis of this is required to sufficiently state whether or not this is the case.

The usability considerations played an integral role in the User Experience Questionnaire (UEQ) results. Tables 13, 14, and 15 displays the processed UEQ data. Examining the scales' mean data shows a noticeable difference between the experimental and control groups scales. Specifically, the scales: *Attractiveness*, *Stimulation*, and *Novelty*. Comparing this to their respective p-value's show that, although close, Attractiveness does not show a significant difference between the two group's mean. Meanwhile, Stimulation and Novelty do show a significant difference between the experimental and control group. As stated in Section 2.3.2.5, Attractiveness is a reference of the basic emotional reaction of liking or disliking something. With a p-value of 0.1034, and the Alpha-Level being 0.05, it is relatively close to provide significant prove that the participant's rather use the full augmented reality integrated version, than the limited one. Why this is the case can be compared to the measures of Stimulation (p-value: 0.0348) and Novelty (p-value: 0.0195). These two scales are of hedonic quality, meaning they appeal to a user's pleasure and avoidance of boredom and discomfort (see Section 2.3.2.5). As a significant difference was found between the two scales and groups, it can be stated that the full AR integrated version pleasing appealed to the users when compared to the limited version. This could be due to the following reasons, 1. Displaying 3D data visualization in augmented reality greatly increases the experience if you are able to move free around the dataset, and 2. The full integrated version was more fun to use than conventional tap gesture manipulations. While this can be derived from the results, the scales of pragmatic quality: *Perspicuity*, *Efficiency*, and *Dependability* do not show any significant difference. As these scales concern the interactions with a product and its usefulness, efficiency, and ease of use, it can be stated that between the two versions no difference in pragmatic quality can be found. This also goes well in-hand with the task performance results, where it is difficulty to differentiate between the two groups (see Tables 16 and 17).

Moreover, the standard deviation in both groups can be considered relatively high, as they all suggests a high level of variability or dispersion in the dataset (see Tables 13 and 14). This correlates with the confidence levels and intervals of each group. This estimate reflects the probability that the mean of each scale would fall between the interval range of values, if the experiment is revisited. As the confidence is rather low, it is difficult to state that the true population was within

the sample size. However, indications show that this could be due to the smaller sample size for the experiment. With a larger sample size, the confidence levels and intervals could become more precise.

Although the experimental group scored their user experience to be somewhat better than the control group's experience, the NASA-TLX task scores show that the experimental group had to use more mental workload to complete the tasks. All tasks, but Task 4, in the experimental group scored a higher NASA-TLX score than the control group (see Tables 20 and 21). Despite the fact that the Stimulation and Novelty scales from the UEQ framework show a significant difference between the two group's means, this could tell the sole reason as to why the experimental group has spent more mental workload while completing the tasks. As the experimental group used the full augmented reality integrated version, they got exposed to a new and unfamiliar way of viewing and analyzing data. This alone could play a part as to why there is a difference in the aggregated NASA-TLX scores. While admitting there are some differences in the scores from one group to another it is difficult to state that a significant difference can be found. As seen in Table 22 the closest tasks to show a significant difference are Tasks 1 (p-value: 0.233), 2 (p-value: 0.203), and 5 (p-value: 0.232). While not being close to the 0.05 benchmark, the task results do show a difference between the two groups, with the experimental group spending more mental workload on task completion. This could be correlated to the time spend on a task. Respectively, the experimental group spend more time on task completion for each of the five tasks. Especially Tasks 2 and 5 showed an extensive difference between the two groups. This could indicate that: 1. The experimental group found their version to be more fun and pleasing to use, and as a result they spend more time in solving the tasks due to the fascination and exploratory aspects of using something new. 2. The experimental version required more mental workload to comprehend, in which the users had to use more time to understand the tasks and datasets. Although the tasks and datasets were the same, the presentation, in this case an iPad with and without a static position, seemed to play a role in a users mental workload. This is the extraneous factor of cognitive load. We wished for the experimental group to show lesser cognitive load levels than the control group, however, this is not the case. This could be due to the familiarity aspect of using a tool like you are used to, rather than a new alternative. Nevertheless, as the results are not significantly different it could suggest, with time, that the full AR integrated version could provide less cognitive strain if used more frequently by the population.

As we did not manage to show a positive result of cognitive load in the experimental group, it could be discussed as to whether or not we used the best fitting framework, or if we had to manipulate it. In terms of manipulation, it could be seen to influence and bias the questionnaire by removing or adding new items, especially to a framework with high reliability. Had we removed or added new items to the questionnaire (UEQ or NASA-TLX), the experiment could have showed a different outcome. However, in doing so, we had to check the "new" questionnaire for reliability on a large sample size, which was found to be an extensive and out of scope task for a study of

this size. Alternatively, we could have weighted the six NASA-TLX sub-scales differently. For this experiment the *Mental Demand* (0.3) and *Effort* (0.3) category had the largest weights by being three times the weight compared to the other sub-scales (see Section 3). This weight was determined based on the sub-scales we found to be the most influential on a users mental workload. In regards to the Mental Demand, we assumed this to directly correlate with the three aspects of cognitive load (intrinsic, extraneous, germane), while we found Effort to include aspects of both the Mental and Physical Demands. Categories like *Temporal Demand* (0.1), *Performance* (0.1), and *Frustration* (0.1) were all found to be less impactful. Although important in the greater picture of cognitive load, for our narrow scope their impact was lessened due to there being no time constraints, no stress on providing a correct answer, and trying to mitigate frustration by having the users complete the tasks in their preferred tempo. Using a different approach could have resulted in a different outcome. If we had based our experimental design of cognitive load on Ouwehand Et al's. four subjective rating scales (see Section 2.3.4.1), we possibly could have found a more clear reflection of a user's underlying simple and complex problem solving capabilities. This could have provided more concrete information on each participant, and as a result provide evidence that the full AR integrated version did in fact prove to be less cognitively straining, than the limited version. However, this must be assessed in a future study to significantly prove this hypothesis.

With a higher success rate for answers across both groups compared to the third iteration, it might be worth considering what the cause of this could be. Even by comparing the third iteration task results to the experimental group, which mostly had the same testing conditions, the results differ from around an average of 92.2% (see Table 17) to 79.8% (see Table 11). We are confident that most of the differences between the conditions, such as the physiological measuring equipment have not played a part in improving the task performances, however by breaking up the tasks by having brief questionnaires in between, as opposed to completing them continuously, could have possibly caused breathing room for participants and additional time to rethink and evaluate between tasks. We are uncertain though, as the case could also as easily have been opposite and caused an interruption in their flow, meaning that the performance disparity would have been caused elsewhere. As tasks were presented identically by the same test facilitator, with the same information between the third iteration and final test, the only logical reason we are confident in labelling as the likely reason, would be the sample size. With the difference of sample size, n = 8 and n = 29, the likely scenario has probably been that a few participants who performed poorly during the third iteration tests were slight outliers in the grand scheme of the target group, and that the case of slightly more wrong answers, weigh much higher when the sample size is not as large.

Based on these task performance results, we can also see that the experimental group had a slightly higher correct results than the control group (see Tables 16 and 17). While this does not inherently influence answering the research question on its own, we might be able to argue that the AR

application inherently is a successful product on its own. This should mainly be understood as a general contemplation and with a grain of salt however, as the comparison is made with an application that has not been validated as a state-of-the-art application, which to the best of its ability reflects as a prime example of a 3D scatter plot visualization tool with the most optimal interactivity. Furthermore, no significance test has confirmed that these differences are due to the difference of AR visualization vs. tap gesture visualization, as opposed to a random difference in the participant's sampled during the test. The results however, do decrease the likelihood of an opposite significance, i.e. the control group having an easier time to answer correctly compared to the experimental group.

The experimental and control groups had very similar levels of experience with data visualization, which give a good distribution that otherwise could have caused one group to outperform the other. However, the experience of data visualization might differ from their experience or knowledge of statistics. This means that some participants might have had harder time understanding the more statistically abstract concepts such as finding three dimensional correlations and/or estimating position of points based off specific, but limited, information.

An important aspect to consider for the task difficulty ratings from the test participants, is the order of presenting the tasks. As we intentionally ordered the tasks to appear in order of approximated ascending difficulty to avoid introducing the harder tasks first, the difficulty rating could have been influenced by having to learn the system and task structure. What this essentially means is that the perceived difficulty of the first task being slightly harder than estimated, which was the case for the third iteration as well, might have been caused by the participants having to learn how to navigate the system. This could include figuring out how to visualize using the axes and viewing angles, understanding the context as a whole, and even understanding the task as well as what type of answer to give: i.e. yes/no, a number, detailed description, etc. Based off this assumption, we do not necessarily conclude that the first task was harder than the second task, but rather that being an effect of the initial learning condition of the prototype. Alternatively, we could have used randomization or counterbalancing, however we expect this would have had an increased spike in difficulty for harder tasks being presented earlier.

One of the measures we also included for the test, was observing an estimate of how much each individual participant moved around. While these values were relatively arbitrary, we ensured that the values were estimated across all participants and under the same conditions, as well as provided by the same test facilitator. This ensured that the values would be as relative to each other as possible. The initial idea for including these measures were to figure out if there were any obvious correlations between the amount of movement and their performance, but as there were no obvious indications and that the values were estimated so arbitrarily, we decided not to go too in-depth with the analysis of this aspect. Instead it might be worth considering a different approach in a future study. One idea would be to track and log the movement data from the device, or alternatively, track the virtual position relative to the 3D plot in the application so we

could send the signals similarly to the other event signals we track during the test. If going in this direction, it would definitely be worth performing a much more in-depth test for correlation between the movement and the task performance, cognitive load, user experience, and/or even usability. There is a possibility that they all could rely on the success being based on how much the participant navigated the 3D space, as opposed to lesser movement which would have limited their perspectives and visualization overall.

The results from Task 5 also have some interesting considerations due to the fact that the control group scored incorrect on this task far more frequently than the experimental group. The Task 5 difficulty is also declared to be much harder for the control group, which is an expected correlation. The reason for this is more unclear however, as one of the only explanations, apart from a statistical deviation by random chance, would be the dynamic version was more efficient for solving this task. Task 5 primarily involved getting an overview of the general tendencies of all the data points and their relationship between all three axes, and by estimating where a data point with given variables would be positioned.

While we consider the participants experience in 3D data visualization, the task of estimating the position of a hypothetical data point might require more general statistical knowledge and experience, which we do not account for in the experimental test. This could have been a factor in the difference between the two groups, however a likely factor could also have been the actual 3D navigation between the two group, i.e. moving around in AR and the ability to easier navigate close to specific points to highlight them. An issue that could have made it more difficult for the static control group would be the tap gesture interactions deselecting highlighted object's when manipulating the 3D plot. Additionally, the experimental group could have had an advantage in zooming in on specific points more accurately in AR than the control group, which also could explain part of the disparity. This could likely be opposite however, as this contradicts the task duration times for Tasks 3 and 5 as they have the largest disparity between the groups in terms of time to complete. This propagates the idea of the disparity being caused by sample randomness rather than anything else, however, we do not disregard the possibility of the dynamic version of the application having improved navigational efficiency. An important aspect to consider with the results of the control group as a whole is the tap gesture feature usability. While this was tested internally, this feature lacked external usability validation. While we assume that this has not made any significant difference overall, through observations during the test, thourough testing of this feature could have circumvented a few issues such as the minor inconvenience of deselecting highlighted objects as an example.

The task that proved to be the hardest of all tasks was Task 4. It involved the participant to distinguish multiple correlations from all three axes at once. We expected this task to be harder than most of the other tasks, just by nature of the task having one more dimension than Task 1 and 2, and thus increased complexity. However, this task might also have moved more into the domain of statistical theory than the other tasks, which have resulted in many of the participants,

even those who rated themselves as very familiar, to fail the task. Through observation most of the wrong answers for Task 4 were due to misinterpreting one of the axes, or trying to divide the data points in their given species clusters, instead of interpreting all the data points as a whole.

The easiest task across the two groups based on the data, was Task 2. The participants had to identify a correlation in the dataset using the two axes: Living Area and Price. While it does make sense as we also estimated this task to be one of the easiest alongside Task 1, this task also has an inherent obviousness to it. For example, without even looking at the scatter plot, a participant could very well intuit that houses with a higher living area also would be more expensive. Through observation during testing however, it was never the case that any participant would disregard the data and just give an answer, but in the grand scheme of the task performance, more participants across both groups could have had an easier time answering this task overall. As previously mentioned, the reason this task was rated easier than the first task, despite them being very similar could also have been influenced by the order of presentation, i.e. having to learn the application and understand the context.

When estimating the task difficulty we decided to have three simple understandable difficulties: Easy, Medium, and Hard, but an alternative could have been to have the difficulty been estimated on i.e. a 1-7 point Likert scale. The reasoning for using the three difficulty categories was to simplify it for the participants, allowing them to easy and quickly give any given task a rating. This does take away from the overall accuracy that a 1-7 or 1-10 point Likert scale would have in a wider range of intervals between Easy and Hard. There might have been some cases where a test participant rated the difficulty one thing but was very close to the threshold of voting a different rating, which would have been less likely using a wider range. It would also have been much easier for us to get a precise estimate across all the participants of how difficult each task was. Perhaps having more specific difficulty ratings could also help answer the disparity between the two groups task difficulty ratings.

Certain challenges occurred during the final evaluation, regarding the Blood Volume Pulse (BVP) sensor. Some of the participants had found the BVP sensor to be somewhat intrusive, particularly the experimental group (EG), who had to move around with and carried the iPad with their hand equipped with the sensors. In order to use the iPad, the participants would normally carry it with their non dominant hand, which was also equipped with the BVP sensor and using their other, more dominant hand for tap gestures and interacting with the iPad. The pressure and weight of the iPad on the BVP sensor, and the increased physical movement required in the experimental version, likely generated more signal noise compared to the control group who who stand still for most of the time. However, the pre-processing used filter and noise reduction to the BVP data beforehand (see Section 3.3.3.1). Other influences, such as environmental/weather conditions could have significantly influenced the BVP data, as sufficient lighting is necessary for the device to measure properly. The lightning in room with windows can vary from day to day and this could have some influence on the BVP data. Although the testing room generally offered good conditions

with minimal static noise and sunlight, extra sunlight in the mornings could have interfered with the BVP data, however, the lightning in the room would remain reasonably static for the remainder of the day. As the experimental group was required to move around physically, it naturally introduced more noise into the BVP signal than the control group, which remained physically static for most of the time. The increased movement from the experimental is important to consider when analyzing the BVP data, as it could affect heart rate measurements but also provide more noise. For future research, implementing more advanced filtering techniques could likely have improved the data accuracy and validity of the data. External factors such as the nervousness, presence of other people (test conductors) and environmental noise could also have influenced the participants heart rate measurements.

Another important consideration is whether or not the baseline HRV measurements should have been subtracted from SDNN and RMSSD values, as with BPM. Adjusting for baseline values could provide a clearer view of the physiological changes induced by the tasks, but finding sources that mentioned this approach, was difficult. The experimental group had higher BPM (Beats Per Minute) across all tasks compared to the control group. This was likely a result from the increased physical activity required in the dynamic AR environment. Since BPM is directly correlated with physical activity, the additional movement in the experimental tasks explain the higher heart rates measured. This is partially evident in Task 2, where the difference between groups was almost statistically significant ($p = 0.0538$). The experimental group also had higher cognitive load levels as reported by the NASA-TLX (see Tables 20 and 21), alongside greater heart rate variability (SDNN and RMSSD), which could indicate a more demanding cognitive environment due to the dynamic AR integration. SDNN values, which reflect overall heart rate variability, showed that both groups maintained values above 50, which indicates acceptable health levels. However, the control group exhibited notably lower SDNN values in Tasks 1, 3, 4, and 5 compared to the experimental group. Lower SDNN in the control group suggests lower stress and cognitive load, consistent with the findings from the NASA-TLX questionnaire (see Tables 20, 21 and 22). None of the differences in SDNN were statistically significant, however Task 1 ($p = 0.0717$) and Task 5 ($p = 0.0620$) were close to the significance level (0.05). External factors such as the presence of other people and environmental noise likely influenced heart rate variability. During testing, at least two test facilitators were present in the room at all times, while the participants also had the physiological equipment equipped to their hands, which for some might be intrusive. On the final day of testing, construction noise was also present, which could have introduced additional stress and affected HRV measurements, however, this would only entail about 10 test participants.

There have previously been some debate on what is the appropriate recording time for HRV measures. While a standard 5-minute recording time is conventional [107] according to some studies, some sources suggest shorter duration (60-240 seconds) may be sufficient for recording SDNN, though RMSSD should be interpreted cautiously if shorter times are used [93][94]. For future research, it would make sense to carefully consider these guides lines beforehand to get the most

accurate and reliable HRV data. It could have proven useful to include subtract baseline HRV measures from the task HRV measures as it can help understand individual stress responses relative to their baseline, enhancing the correlation analysis between cognitive load and HRV. Subtracting the baseline from the task data would have helped giving more more context to the HRV measures and help distinguish between stress that is the result of a task from individual baseline variations. Although we were in possession of the baseline HRV (SDNN and RMSSD) values before tasks, they were not subtracted or normalized from the HRV task values. Subtracting the baseline and normalizing the percentage of change for HRV measures (SDNN and RMSSD), like with the BPM measure, would have been beneficial. However, it was unsure whether this approach would be viable in our study due to the lack of supporting sources, so we decided to use the HRV measures without normalization or subtracting the baseline. To gain a better understanding of the participants cognitive load levels, additional measures such as respiration rate (RR) or blood oxygen saturation (SpO2) could have been included in the testing setup, as these measures have been found to correlate with cognitive load [108][109]. Additionally these measures, can be derived directly from the BVP data and would have provided us with further measures and insights into the cognitive and physical demands of the tasks and two versions.

Overall, the BVP data analysis provides us with valuable insights into the physiological responses of participants, showing increased cognitive and physical demands of the dynamic AR environment, which also correlates with the NASA-TLX results (see Tables 20 and 21. However, improving environmental conditions, reduction of movement related noise, baseline adjustments for HRV values, and recording times could have enhanced the validity and reliability of these findings.

The Electrodermal Activity (EDA) data (see Table 24) collected in this study were influenced by various environmental and participant-specific factors. Factors such as ambient temperature, humidity, and participants hydration levels could have influenced the EDA data. No measures were taken to control these factors, such as using a hygrometer to monitor room temperature and humidity. In hindsight, this could have provided more controlled and reliable EDA data. The EDA electrodes were positioned on participants fingers by utilizing the adhesive surfaces on the electrodes. In most cases, these electrodes were tightly attached and did not cause any issues throughout the testing sessions, however, a few participants with very sweaty hands caused issues as the electrodes were peeling off during the testing session, which forced the test facilitator to properly reattach them. It should be noted that this issue was only relevant for about 2-3 test participants. This was likely due to individual participants conditions and not environmental factors, as the participants who experienced this issue also had a very high raw EDA mean. The raw EDA mean values showed a higher percentage increase in the experimental group, except for Task 5. The experimental group also achieved significantly higher numbers of SCR peaks. The Mann-Whitney U Test indicated that these differences were significant in 4 out of 5 tasks, with the remaining task (Task 1) also being close ($p = 0.0688$). This suggests greater emotional arousal and cognitive load in the dynamic AR environment. The increased SCR peaks in the experimental group might

be because of the dynamic AR application being more engaging and mentally stimulating. The spatial nature of the dynamic version allowed participants to move around, potentially increasing engagement and immersion. Additionally, participants frequently expressed enthusiasm and amazement over the dynamic version compared to the static version. It is important to note that the number of SCR peaks was normalized by the duration of each task to account for varying completion times, which should take into account the additional time spent by the experimental group participants. Having both higher SCR peaks and amplitudes suggest that the experimental group experienced more intense and frequent emotional responses, which might reflect higher cognitive load and stress levels during the tasks. In this context, it could have been interesting to allow participants to interact with the dynamic AR solution across multiple sessions, assessing whether their physiological responses and cognitive load diminished as they adapted more and more to the technology and spatial elements. While the study primarily focused on skin conductance responses (SCR), additional measures such as skin conductance levels (SCL) could have provided more insights. SCL measures longer-term arousal but requires extended recording times than SCR [100]. Including SCL in future studies could offer a more comprehensive understanding of both immediate and sustained emotional responses, but would likely require longer test duration's to be valid.

Generally, the higher EDA measures in the experimental group across all EDA measures, align with the higher NASA-TLX scores and blood volume pulse (BVP) data, suggesting a consistent relationship between physiological emotional/physical responses and self-reported cognitive load. It should be noted that since the dynamic AR environment involved more physical movement, the impact of physical activity on EDA responses must be considered when using this data. Higher physical activity levels might increase EDA measures, reflecting not only cognitive load but also physical workload. Overall, the EDA data provided valuable insights into participants emotional arousal and cognitive load during AR tasks. The correlations with NASA-TLX and BVP metrics support the validity of using EDA as a measure of cognitive and emotional responses in AR environments.

Retrospectively, it was mentioned in the analysis that Bach Et al. concluded in their study, that they found mobile devices were easy-to-use and had great portability but lacked behind in speed and precision compared to desktop/static environments [20]. Our findings somewhat confirm this, as the dynamic group had a lot higher task duration across all tasks compared to those who used to static version which included more familiar ways of interacting with the data, compared to dynamic version's spatial approach. The study by Bach Et al. also suggested the potential for increased cognitive load in AR environments, which our findings also seems to confirm.

An important takeaway from this is the need to balance the highly immersive and engaging qualities of the spatial AR application with the increased cognitive load implications of AR and VR. This balance must be carefully considered in the development of these types of projects to ensure that the benefits of immersion do not come at the cost of decreased user experience or efficiency.

9 Conclusion

In conclusion, our study aimed to address the two following questions:

Research question 1: *How does an augmented reality application for data visualization enhance the user experience compared to a static version?*

Research question 2: *By use of a multimodal cognitive load measuring approach, how does an augmented reality application for data visualization influence extraneous cognitive load?*

Through our research, we found that enhancing user experience in an augmented reality 3D data visualization application was a complex task. Using the User Experience Questionnaire (UEQ) helped to significantly prove that the sub-scales, *Stimulation* and *Novelty*, did in fact show a significant difference between the experimental and control group's mean values (see Table 15). These two sub-scales, along with *Attractiveness* almost reaching a significant benchmark, are related to the hedonic qualities as opposed to the application's pragmatic qualities like efficiency and ease of use, which is implied by the dynamic application's more engaging and immersive nature. This together with the task performance results suggest that the user experience was generally better in the experimental group, when compared to the control group. Consequently, we can accept the alternative hypothesis that: data visualization in augmented reality enhances the user experience when compared to a static version (see Section 2.4).

Through our collection of physiological data, task performance measures, and the NASA-TLX questionnaire, it was found that the experimental group achieved higher cognitive load levels across most of the tasks and measures. Specifically, the NASA-TLX self-report questionnaire indicated higher cognitive load levels for the experimental group, although not statistically significant when compared to the control group (see Table 22). Additionally, the experimental group had increased levels of Electrodermal Activity (EDA) across all measures and tasks, particularly a big increase in SCR Peaks and SCR Peak Amplitudes, which the Mann-Whitney U Test also proved to be statistically significant in favor of the experimental group (see Table 24). We conclude, that this likely indicates increased emotional arousal and cognitive load levels in the dynamic AR application, which also aligns with the NASA-TLX questionnaire (see Tables 20 and 21), but we also acknowledge that this could be caused by the increased physical movement that it required, as well as the additional signal noise caused by this movement. This is furthermore confirmed by the Blood Volume Pulse (BVP) data, which although not statistically significant, showed that the dynamic AR application had higher physical demands, as indicated by the Beats Per Minute (BPM) data (see Table 23). On the other hand, the experimental group had an overall increased heart rate variability (HRV) across most tasks, although this should be approached sceptically as the HRV results were not statistically significant. This likely suggest that they experienced greater physiological arousal and maintained better autonomic regulation during the tasks. The results of the physiological data (EDA and BVP) also correlate with the task performance metric, specifically task duration, where the experimental group spent 20-30% more time completing each task

on average. The increase in time duration might also be caused by the increased Stimulation and Novelty as shown by the UEQ results (see Table 15). Collectively, these measures help to accept the alternative hypothesis that: an augmented reality application for data visualization has increased extraneous cognitive load when compared to a static version. However, there is ample room for further discussion and future work to substantially enhance the validity and reliability of the results.

References

[1] Abdullah M. Al-Ansi et al. "Analyzing augmented reality (AR) and virtual reality (VR) recent development in education". In: *Social Sciences & Humanities Open* 8.1 (2023), p. 100532. ISSN: 2590-2911. DOI: https://doi.org/10.1016/j.ssaho.2023.100532. URL: https://www.sciencedirect.com/science/article/pii/S2590291123001377.

[2] Julie Carmigniani and Borko Furht. "Augmented reality: an overview". In: *Handbook of augmented reality* (2011), pp. 3–46.

[3] GameSpot. *Pokemon Go Is Getting A Better AR Mode*. https://www.gamespot.com/articles/pokemon-go-is-getting-a-better-ar-mode/1100-6455762/. Accessed: (February 13, 2024). 2017.

[4] Ivan E Sutherland. "A head-mounted three dimensional display". In: *Proceedings of the December 9-11, 1968, fall joint computer conference, part I*. 1968, pp. 757–764.

[5] Niantic. *Pokemon GO*. https://pokemongolive.com/. Accessed: (February 11, 2024). 2024.

[6] Xiuquan Qiao et al. "Web AR: A promising future for mobile augmented reality—State of the art, challenges, and insights". In: *Proceedings of the IEEE* 107.4 (2019), pp. 651–666.

[7] Nintendo. *Mario Kart Live: Home Circuit*. https://www.nintendo.com/us/store/products/mario-kart-live-home-circuit-switch/. Accessed: (3th May, 2024). 2020.

[8] Lego. *Welcome to the Hidden Side*. https://www.lego.com/en-dk/product/welcome-to-the-hidden-side-70427. Accessed: (3th May, 2024). 2019.

[9] IKEA. *IKEA Place app launched to help people virtually place furniture at home*. https://www.ikea.com/global/en/newsroom/innovation/ikea-launches-ikea-place-a-new-app-that-allows-people-to-virtually-place-furniture-in-their-home-170912/. Accessed: (3th May, 2024). 2018.

[10] Paul Milgram et al. "Augmented reality: A class of displays on the reality-virtuality continuum". In: *Telemanipulator and telepresence technologies*. Vol. 2351. Spie. 1995, pp. 282–292.

[11] Richard Skarbez, Missie Smith, and Mary C Whitton. "Revisiting Milgram and Kishino's reality-virtuality continuum". In: *Frontiers in Virtual Reality* 2 (2021), p. 647997.

[12] Ulla Wandinger. "Introduction to lidar". In: *Lidar: range-resolved optical remote sensing of the atmosphere*. Springer, 2005, pp. 1–18.

[13] LLC. 8th Wall. *8th Wall*. https://www.8thwall.com/. Accessed: (February 11, 2024). 2024.

[14] Mathias A. W. Kristiansen et al. "Dynamic Difficulty Adjustment Using EEG Data: Measuring Engagement In An Eco-Game". In: *AAU MED7 - Semester Project* (2022).

[15] Rasmus K. Schrøder, Stefan L. Olsson, and Valdemar B. Petersen. "Kingstone: Exploring Narrative Closure in an Emergent Narrative". In: *AAU MED8 - Semester Project* (2023).

[16] Unity Technologies. *Unity3D*. https://unity.com/. Accessed: (February 11, 2024). 2024.

[17] Vuforia. *Vuforia*. https://developer.vuforia.com/. Accessed: (February 09, 2024). 2024.

[18] Unity. *AR Foundation*. https://docs.unity3d.com/Packages/com.unity.xr.arfoundation@5.1/manual/index.html. Accessed: (February 11, 2024). 2023.

[19] Tim Mowrer Todd Stinson. *AR Foundation support for ARKit 4 Depth*. https://blog.unity.com/engine-platform/ar-foundation-support-for-arkit-4-depth. Accessed: (February 11, 2024). 2020.

[20] Benjamin Bach et al. "The hologram in my hand: How effective is interactive exploration of 3D visualizations in immersive tangible augmented reality?" In: *IEEE transactions on visualization and computer graphics* 24.1 (2017), pp. 457–467.

[21] Wikipedia. *6 Degrees of Freedom*. https://en.wikipedia.org/wiki/Six_degrees_of_freedom. Accessed: (February 23, 2024). 2023.

[22] Michael Friendly and Howard Wainer. *A history of data visualization and graphic communication*. Harvard University press, 2021.

[23] Lindy Ryan. *The visual imperative: Creating a visual culture of Data Discovery*. 1st. Vol. 9. Elsevier, MK, Morgan Kaufmann, 2016.

[24] Margaret Anne Defeyter, Riccardo Russo, and Pamela Louise McPartlin. "The picture superiority effect in recognition memory: A developmental study using the Response Signal procedure". In: *Cognitive Development* 24.3 (June 2009), pp. 265–273. DOI: 10.1016/j.cogdev.2009.05.002.

[25] Vizzu. *40 Types of Data Visualization Charts and Graphs*. 2024. URL: https://www.vizzu.io/blog/data-visualization-types.

[26] Edward R. Tufte. *The visual display of quantitative information*. Graphics Press, 2007.

[27] Stephen Few. *Show me the numbers: Designing tables and graphs to enlighten*. Analytic Press, 2012.

[28] M Shukla. *The rise of 3D graphics in web design: Techniques and tools*. 2024. URL: https://manoj-shu100.medium.com/the-rise-of-3d-graphics-in-web-design-techniques-and-tools-fd5e05ef3084.

[29] Claus O. Wilke. *Don't go 3D*. Accessed: 2024-05-22. 2020. URL: https://clauswilke.com/dataviz/no-3d.html.

[30] Microsoft. *PowerBI*. 2024. URL: https://powerbi.microsoft.com/.

[31] Travis E. Oliphant et al. *NumPy Reference*. 2024. URL: https://numpy.org/.

[32] Wes McKinney et al. *Pandas: Powerful Python Data Analysis Toolkit*. 2024. URL: https://pandas.pydata.org/.

[33] John D. Hunter et al. *Matplotlib: Visualization with Python*. 2024. URL: https://matplotlib.org/.

[34] Salesforce. *Tableau*. 2024. URL: https://www.tableau.com/.

[35] Michael Waskom et al. *Seaborn: Statistical Data Visualization*. 2024. URL: https://seaborn.pydata.org/.

[36] Plotly Technologies Inc. *Plotly: The front end for ML and data science models*. 2024. URL: https://plotly.com/.

[37] Michael Bostock et al. *D3.js: Data-Driven Documents*. 2024. URL: https://d3js.org/.

[38] Unity Technologies. *Unity VisualLive*. 2024. URL: https://unity.com/products/visuallive.

[39] Unity Technologies. *Unity Industry*. 2024. URL: https://unity.com/solutions/industry.

[40] Unity Technologies. *Unity MARS: Mixed and Augmented Reality Studio*. 2024. URL: https://unity.com/products/unity-mars.

[41] John Smith et al. "Comprehensive Review of Visualization Techniques for Scientific Data". In: *ACM Computing Surveys* 53.1 (2021), pp. 1–38. DOI: 10.1145/3411764.3445593. URL: https://dl.acm.org/doi/fullHtml/10.1145/3411764.3445593.

[42] Torben Schinke, Niels Henze, and Susanne Boll. "Visualization of off-screen objects in mobile augmented reality". In: *Proceedings of the 12th international conference on Human computer interaction with mobile devices and services*. 2010, pp. 313–316.

[43] Andreas Jakl et al. "Enlightening patients with augmented reality". In: *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE. 2020, pp. 195–203.

[44] Donald A. Norman. *The design of everyday things: Revised and expanded edition*. Basic books, 2013.

[45] Donald A. Norman and Stephen W. Draper. "User centered system design". In: *New perspectives on human-computer interaction* (1986).

[46] Donald A. Norman. *The psychology of everyday things.* Basic books, 1988.

[47] Interaction Design Foundation. *User Centered Design*. https://www.interaction-design.org/literature/topics/user-centered-design. Accessed: (20th February, 2023).

[48] Bill Albert and Tom Tullis. *Measuring the User Experience: Collecting, Analyzing, and Presenting UX Metrics*. Morgan Kaufmann, 2022.

[49] Barbara DiCicco-Bloom and Benjamin F Crabtree. "The qualitative research interview". In: *Medical education* 40.4 (2006), pp. 314–321.

[50] Mary Elizabeth Raven and Alicia Flanders. "Using contextual inquiry to learn about your audiences". In: *ACM SIGDOC Asterisk Journal of Computer Documentation* 20.1 (1996), pp. 1–13.

[51] Bettina Laugwitz, Theo Held, and Martin Schrepp. "Construction and evaluation of a user experience questionnaire". In: *HCI and Usability for Education and Work: 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society, USAB 2008, Graz, Austria, November 20-21, 2008. Proceedings 4*. Springer. 2008, pp. 63–76.

[52] Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. "Design and evaluation of a short version of the user experience questionnaire (UEQ-S)". In: *International Journal of Interactive Multimedia and Artificial Intelligence, 4 (6), 103-108.* (2017).

[53] J. Nielsen. *Usability engineering*. Morgan Kaufmann, 1994.

[54] Carol M. Barnum. *Usability Testing Essentials: Ready, set... test!* Elsevier, 2021.

[55] John Brooke. "SUS: A quick and dirty usability scale". In: *Usability Eval. Ind.* 189 (Nov. 1995).

[56] John Sweller. "Cognitive load theory". In: *Psychology of learning and motivation*. Vol. 55. Elsevier, 2011, pp. 37–76.

[57] John Sweller. "Element interactivity and intrinsic, extraneous, and germane cognitive load". In: *Educational psychology review* 22 (2010), pp. 123–138.

[58] David C Geary. "Principles of evolutionary educational psychology". In: *Learning and individual differences* 12.4 (2002), pp. 317–345.

[59] David C Geary and Daniel B Berch. "Evolution and children's cognitive and academic development". In: *Evolutionary perspectives on child development and education* (2016), pp. 217–249.

[60] John Sweller. "Cognitive load theory and educational technology". In: *Educational Technology Research and Development* 68.1 (2020), pp. 1–16.

[61] Fang Chen et al. *Robust multimodal cognitive load measurement*. Springer, 2016.

[62] John Sweller et al. "Measuring cognitive load". In: *Cognitive load theory* (2011), pp. 71–85.

[63] Fred GWC Paas. "Training strategies for attaining transfer of problem-solving skill in statistics: a cognitive-load approach." In: *Journal of educational psychology* 84.4 (1992), p. 429.

[64] Gabriele Cierniak, Katharina Scheiter, and Peter Gerjets. "Explaining the split-attention effect: Is the reduction of extraneous cognitive load accompanied by an increase in germane cognitive load?" In: *Computers in Human Behavior* 25.2 (2009), pp. 315–324.

[65] Jimmie Leppink et al. "Development of an instrument for measuring different types of cognitive load". In: *Behavior research methods* 45 (2013), pp. 1058–1072.

[66] Fred GWC Paas and Jeroen JG Van Merriënboer. "Instructional control of cognitive load in the training of complex cognitive tasks". In: *Educational psychology review* 6 (1994), pp. 351–371.

[67] Fang Chen et al. "Multimodal behavior and interaction as indicators of cognitive load". In: *ACM Transactions on Interactive Intelligent Systems (TiiS)* 2.4 (2013), pp. 1–36.

[68] Sandra G Hart and Lowell E Staveland. "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research". In: *Advances in psychology*. Vol. 52. Elsevier, 1988, pp. 139–183.

[69] Sandra G Hart. "NASA-task load index (NASA-TLX); 20 years later". In: *Proceedings of the human factors and ergonomics society annual meeting*. Vol. 50. 9. Sage publications Sage CA: Los Angeles, CA. 2006, pp. 904–908.

[70] Fred GWC Paas, Jeroen JG Van Merriënboer, and Jos J Adam. "Measurement of cognitive load in instructional research". In: *Perceptual and motor skills* 79.1 (1994), pp. 419–430.

[71] Kim Ouwehand et al. "Measuring cognitive load: Are there more valid alternatives to Likert rating scales?" In: *Frontiers in Education*. Vol. 6. Frontiers Media SA. 2021, p. 702616.

[72] Tamara Van Gog and Fred Paas. "Instructional efficiency: Revisiting the original construct in educational research". In: *Educational psychologist* 43.1 (2008), pp. 16–26.

[73] Christian Gütl et al. "Adele (adaptive e-learning with eye-tracking): Theoretical background, system architecture and application scenarios". In: *European Journal of Open, Distance and E-Learning* 8.2 (2005).

[74] Rasmus K. Schrøder, Stefan L. Olsson, and Valdemar B. Petersen. "Exploring EEG-Based Dynamic Difficulty Adjustment for Enhanced Player Engagement in Video Game". In: *AAU MED8 - MIMSC Mini Project* (2023).

[75] Yuko Suzuki, Fridolin Wild, and Eileen Scanlon. "Measuring cognitive load in augmented reality with physiological methods: A systematic review". In: *Journal of Computer Assisted Learning* (Oct. 2023). DOI: 10.1111/jcal.12882.

[76] Emma J. Nilsson et al. "Let complexity bring clarity: A multidimensional assessment of cognitive load using physiological measures". In: *Frontiers in Neuroergonomics* 3 (Feb. 2022). DOI: 10.3389/fnrgo.2022.787295.

[77] Michael Richter and Kate Slade. "Interpretation of physiological indicators of motivation: Caveats and recommendations". In: *International Journal of Psychophysiology* 119 (2017). The Psychophysiology of Motivation: Body and Brain in Action, pp. 4–10. ISSN: 0167-8760. DOI: https://doi.org/10.1016/j.ijpsycho.2017.04.007. URL: https://www.sciencedirect.com/science/article/pii/S0167876017302532.

[78] Stephanie R. Fishel, Eric R. Muth, and Adam W. Hoover. "Establishing appropriate physiological baseline procedures for real-time physiological measurement". In: *Journal of Cognitive Engineering and Decision Making* 1.3 (Sept. 2007), pp. 286–308. DOI: 10.1518/155534307x255636.

[79] Kwang Bok Kim and Hyun Jae Baek. "Photoplethysmography in Wearable Devices: A Comprehensive Review of Technological Advances, Current Challenges, and Future Directions". In: *Electronics* 12.13 (2023). ISSN: 2079-9292. DOI: 10.3390/electronics12132923. URL: https://www.mdpi.com/2079-9292/12/13/2923.

[80] Swagata Devi and Soumik Roy. "Physiological measurement platform using wireless network with Android application". In: *Informatics in Medicine Unlocked* 7 (2017), pp. 1–13. ISSN: 2352-9148. DOI: https://doi.org/10.1016/j.imu.2017.02.001. URL: https://www.sciencedirect.com/science/article/pii/S2352914817300114.

[81] Mikolaj Buchwald et al. "Electrodermal activity as a measure of cognitive load: A methodological approach". In: *2019 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)* (Sept. 2019). DOI: 10.23919/spa.2019.8936745.

[82] Arne Seeliger, Long Cheng, and Torbjørn Netland. "Augmented reality for Industrial Quality Inspection: An Experiment Assessing Task Performance and Human Factors". In: *Computers in Industry* 151 (Oct. 2023), p. 103985. DOI: 10.1016/j.compind.2023.103985.

[83] Pennsylvania State University. *6.2 - Significance Levels*. https://online.stat.psu.edu/stat200/book/export/html/157. Accessed: 2024-05-13.

[84] Laerd Statistics. *Hypothesis Testing - Significance levels and rejecting or accepting the null hypothesis*. https://statistics.laerd.com/statistical-guides/hypothesis-testing.php. Accessed: 2024-05-13.

[85] S.A. PLUX – Wireless Biosignals. *biosignalsplux Explorer User Manual*. http://biosignalsplux.com/. Accessed: 2024-05-13. 2020.

[86] Christian Kothe. *Lab Streaming Layer (LSL)*. https://github.com/sccn/labstreaminglayer. Accessed: 2024-05-13. 2014.

[87] PluxBioSignal. *BioSignalPlux: HYBRID-8*. 2024. URL: https://www.pluxbiosignals.com/products/hybrid-8.

[88] PluxBioSignal. *BioSignalPlux: All your bigsignals in a single hub*. 2024. URL: https://www.pluxbiosignals.com/pages/biosignalsplux.

[89] S.A. PLUX – Wireless Biosignals. *Blood Volume Pulse (BVP) Sensor Data Sheet*. http://biosignalsplux.com/datasheets/BVP_Sensor_Datasheet.pdf. Accessed: 2024-05-13. 2015.

[90] Paul van Gent. *HeartPy: Python Toolkit for Heart Rate Analysis*. https://github.com/paulvangentcom/heartrate_analysis_python. Accessed: 2024-05-13. 2018.

[91] Fred Shaffer and John P. Ginsberg. "An Overview of Heart Rate Variability Metrics and Norms". In: *Frontiers in Public Health* 5 (2017), p. 258. DOI: 10.3389/fpubh.2017.00258.

[92] Task Force of the European Society of Cardiology, the North American Society of Pacing, and Electrophysiology. "Heart Rate Variability: Standards of Measurement, Physiological Interpretation, and Clinical Use". In: *Circulation* 93 (1996), pp. 1043–1065. DOI: 10.1161/01.CIR.93.5.1043.

[93] R. M. Kleiger et al. "Decreased heart rate variability and its association with increased mortality after acute myocardial infarction". In: *American Journal of Cardiology* 59.4 (1987), pp. 256–262. DOI: 10.1016/0002-9149(87)90795-8.

[94] Fred Shaffer and John P. Ginsberg. "An Overview of Heart Rate Variability Metrics and Norms". In: *Frontiers in Public Health* 5 (2017), p. 258. DOI: 10.3389/fpubh.2017.00258.

[95] Hugo F. Posada-Quintero and Ki H. Chon. "Innovations in Electrodermal Activity Data Collection and Signal Processing: A Systematic Review". In: *Sensors* 20.2 (2020). Accessed: 2024-05-15, p. 479. URL: https://doi.org/10.3390/s20020479.

[96] M. E. Dawson, A. M. Schell, and D. L. Filion. "The Influence of Humidity on Electrodermal Response". In: *Handbook of Psychophysiology* (1990). Cambridge University Press, pp. 200–223.

[97] H. R. Lieberman. "Hydration and Human Performance". In: *Journal of the American College of Nutrition* 26 (2007), 452S–458S. DOI: 10.1080/07315724.2007.10719656.

[98] Wolfram Boucsein. *Electrodermal Activity*. Springer, 2012. DOI: 10.1007/978-1-4614-1126-0.

[99] Dorus J van der Mee et al. "Validity of electrodermal activity-based measures of sympathetic nervous system activity from a wrist-worn device". In: *International Journal of Psychophysiology* 168 (2021), pp. 52–64. DOI: 10.1016/j.ijpsycho.2021.08.003.

[100] EDA Guidelines. *Peak Detection Approach*. Accessed: 2024-05-15. 2024. URL: https://edaguidelines.github.io/analysis/pda.

[101] S.A. PLUX – Wireless Biosignals. *Electrodermal Activity (EDA) Sensor Data Sheet*. http://biosignalsplux.com/datasheets/EDA_Sensor_Datasheet.pdf. Accessed: 2024-05-13. 2020.

[102] Dominique Makowski and et al. *NeuroKit2: A Python Toolbox for Neurophysiological Signal Processing*. https://neuropsychology.github.io/NeuroKit/functions/eda.html. Accessed: 2024-05-13. 2021.

[103] Michael Winter et al. "Towards the Applicability of Measuring the Electrodermal Activity in the Context of Process Model Comprehension: Feasibility Study". In: *Sensors* 20 (Aug. 2020), p. 4561. DOI: 10.3390/s20164561.

[104] Mathias Benedek and Christian Kaernbach. "A continuous measure of phasic electrodermal activity". In: *Journal of Neuroscience Methods* 190.1 (2010), pp. 80–91. DOI: 10.1016/j.jneumeth.2010.04.028.

[105] Carlos Wilkes. *Lean Touch | Unity Asset Store*. https://assetstore.unity.com/packages/tools/input-management/lean-touch-30111. Accessed: 2024-03-13. 2024.

[106] Carlos Wilkes. *Lean Touch Documentation*. https://carloswilkes.com/Documentation/LeanTouch. Accessed: 2024-03-13. 2024.

[107] Kyoung Min et al. "Is 5-Minute Heart Rate Variability a Useful Measure for Monitoring the Autonomic Nervous System of Workers?" In: *International heart journal* 49 (Apr. 2008), pp. 175–81. DOI: 10.1536/ihj.49.175.

[108] Author(s). "Title of the IEEE Paper". In: *IEEE Transactions on [specific field]* Vol. XX.No. YY (2022), pp. 1–8. DOI: 10.1109/XXXXX.2022.9851405. URL: https://ieeexplore.ieee.org/document/9851405.

[109] Bryn Farnsworth. *How Respiration Affects the Brain*. Accessed: 2024-05-19. 2024. URL: https://imotions.com/blog/insights/how-respiration-affects-the-brain/.

10 Appendix

10.1 Appendix A

Introduction:
We are a master student group who want to develop an application for AR data visualization. This interview is part of our preliminary work, where we investigate people's preferences and knowledge in statistics. Feel free to let us know when to elaborate on any of the questions, or if you have any other questions, feel free to ask.

Q1. What level would you rate your statistical knowledge?
Fundamental Level (Descriptive statistics, Basic probability concepts, Hypothesis testing)
Intermediate Level (Regression analysis, Experimental design, Multivariate analysis)
Advanced Level (Time series analysis, Bayesian statistics, Machine learning algorithms)

Q2. Can you describe your current process for visualizing and analyzing data?
What tools or software do you typically use?

Q3. Which datasets do you find yourself using most often?
Scatter plots
Line graphs (or Line charts)
Bar charts (or Bar graphs)
Histograms
Pie charts
Box plots
Heat maps
Others?

Q4. Can you discuss any challenges or difficulties you face when having to visualize data?
How do you typically address these challenges?
Visual aesthetics (color, shapes, sizes)
Visualizing data with more than three dimensions?
Visual bias? Perspective of size etc.
Visualizing all the data at once? E.g. Clutter and Occlusion

Q4.1 Do you think interactivity and real-time manipulation would help in data visualization?
Highlight, change/scale axis, colors etc.

Q5. Have you ever used augmented reality (AR) technology before?
If yes, what was your experience like?
What part of AR did you enjoy?
What part of AR did you not enjoy?
Would you consider using Augmented Reality for visualizing and manipulating your data?

Q6. How do you think AR technology could enhance your current data visualization workflows?
A more engaging or immersive experience?

Q6.1. In the context of 3D data; Do you think it could be more intuitive to understand by moving around a mobile camera than a traditional mouse/keyboard to rotate and zoom?

Q7. How would you measure the success or effectiveness of an AR data visualization tool?
Easy to use?
Easy to learn?
Easy to visualize all data at once?
Compatibility with different types of data sets?
Complexity or simplicity when it comes to features?

Q7.1 How would you prioritize the mentioned measures from most important to least?

Q8. Would it be helpful to use data visualization in AR as a presentation/explanation medium.

10.2 Appendix B

10.2.1 Participant 1: Fundamental Level

Interviewer: "For a brief introduction, we are a master student group who want to develop an application for AR data visualization. This interview is part of our preliminary work, where we investigate people's preferences and knowledge in statistics. If you have any questions, or want us to elaborate on something, please let us know - and feel free to ask any questions you might have."

"So, my first question to you is - What level would you rate your statistical knowledge be? So, we have three points, if you use any of the methodologies in a category you will be considered to fit that level group."

Respondent: "Well, since we're pretty early in Medialogy, it is pretty much a fundamental level, I think."

Interviewer: "Can you describe your current process for visualizing and analyzing data?"

Respondent: "Oh, uhm, right now we're using Python to work with a lot of data, but personally I have also used a lot of Excel, and it is usually just simple graphs with data points put on the x and y-axis. So, it is nothing more complicated than that."

Interviewer: "This kind of leads me to the next question - Which type of datasets do you find yourself using most often?"

Respondent: "Hmm, usually with two variables *mumbles* in a graph."

Interviewer: "Could it be histograms, pie charts, box plots?"

Respondent: Definitely, histograms. We haven't used any box plots yet, but I personally think they are pretty useful, for later. But right now, we have just utilized histograms and not really anything else."

Interviewer: "So, you use percentages?"

Respondent: "Yeah."

Interviewer: "Okay, this might be a little difficult to answer, but please try your best. Can you discuss any challenges or difficulties you face when having to visualize data?

Respondent: "Hmm, I think one of the most obnoxious things is when you get some error in your data you kind of have to sort out. For example, if you try to calculate something and you get a completely extreme value that doesn't make any sense for your research beforehand or anything of that source. I think that is one of the biggest problems, and then sometimes like when you make something in Python it just might not look the way you wanted it to look, and then you have to go in and try to figure out why that is the case. Uhm, but mostly... Things might not go the way you expecting, and that can sort of be discouraging and make it just harder to work with."

Interviewer: "And when you work with these challenges - How do you typically address these challenges?"

Respondent: "I don't know. I guess you just kind of work with them until they are better. There is nothing else than that. You take it front face and then depending... it is a lot case by case basis i would say. It is not as easy to come with a specific method that I would utilize for getting rid of the problems."

Interviewer: "Have you experienced problems like visual aesthetics of colors, shapes, and size - maybe some cluttering?"

Respondent: "Yeah, there was... We made a histograms recently in Python. I remember that all the data looked like one big smush, because there wasn't gaps between the actual data. So, it was hard to see what was what. It was hard to see how many people that said 20, and how many said 21. It just looked like the same.

Interviewer: "Do you think interactivity and real-time manipulation could help you in visualizing your data?"

Respondent: "I think it would be pretty good if I could just change certain things about how I want it, and then as you said, in real-time see how it, the visualization, changes, because then I can find, okay, exactly when does it click. So, I can gradually see the changes of my data. Uhm, of the visualization of the data. So, I can gradually see how it changes, so I can find exactly what I want to show to others. So, I think that is a great idea."

Interviewer: "And this leads me to kind of wanting to show it to others. Have you ever used augmented reality before?"

Respondent: "No."

Interviewer: "Okay. You have never used Pokémon Go?"

Respondent: "Oh! Used. I was thinking... But yeah, I have used Pokémon Go, so yes I have."

Interviewer: "Okay, so you do know what augmented reality is?"

Respondent: "Yes."

Interviewer: "Would you consider using augmented reality for visualizing and manipulating your data?"

Respondent: "I don't... I can't currently see how you would do that, but I wouldn't be completely against it. It's kind of an unknown area for me. So, if somebody has a great way to do it, then I would say that would be great. But I can't visualize how it would actually function and how it would work.

Interviewer: "Okay, that's is fair enough. How do you think augmented reality could enhance your current data visualization workflow? So, you mentioned you thought that the interactivity and real-time manipulation could help. Do you think that if you put AR on top of that it would help you visualize your workflow?"

Respondent: "Hmm. I still can't really see how it would actually function other than being on a screen. If you have like... If you interpret the data differently using augmented reality, I guess you could like see more in spectrum of how... feel in real life, instead of just on a computer screen because that is just a square that is on the computer screen. But I don't really. I guess an idea would be, if you had to make a histogram about heights. You could use augmented reality and also see how high it would be in real life, so it would also give you more perspective, I think - on the data. So, I guess yeah, it would give more of an in-depth look at data."

Interviewer: "Yeah, makes sense. So in the context of 3D data. Do you think it could be more intuitive to understand the data by moving around a mobile camera than a traditional mouse/keyboard input?"

Respondent: "I would probably say because not everyone would know how to navigate a 3D space on a computer. But just because we live in a 3D world and all you have to do is to pull of your phone or whatever, or how it gets implemented. It makes a lot closer to home, so I think it would be a lot more... it would be closer to home, so everybody would have easier to understand it. They would immediately be able to know, okay if I walk around I can see it this way or I can look at it from this angle. When you are talking about 3D."

Interviewer: "Do you think it would be more engaging or immersive?"

Respondent: "Yeah."

Interviewer: "So, you were to measure the success or effectiveness of an augmented reality visualization tool - what would the measurements be - would it be that it is easy to use?

Respondent: "Hmm, yeah, I would definitely say the ease of use and also how easy it is to understand the data. Maybe in some broader context. Because one thing is just looking at a bunch of numbers and one thing is looking at a histogram, but understanding it - like what are these numbers and this data actually mean in a broader context is where it kind of gets hard, and I think a lot of people can fall off. So, yeah how easy it is to use and how it more visualize the data and how more easily people understand it."

Interviewer: "Also the learnability?"

Respondent: "Yeah."

Interviewer: "What about the complexity and simplicity when it comes to specific features? If you compare it to, you mentioned Excel and Python, do you think that an AR data visualization tool could make it more complex or simple?

Respondent: "Uhm, i think that depends on how easy it is to get into it and use it. Because obviously with stuff Excel it is pretty accessible you can pretty much just go into it and just put numbers down, but there is so much things that the program can do. So, I think it would make it easier if the AR experience would be easy to get into, easy to start using, but also has the possibility of being pretty complex, so you would find a reason to use it, instead of just, okay I will go back to my Excel or Python."

Interviewer: "So, do you think it would be helpful to use data visualization in AR as a presentation or explanation medium? Let's say you have some data, you want to present it to others, and you want to explain it to others, do you think showing your data as an AR visualization would be better than just showing a 2D or 3D plot?"

Respondent: "I don't know. Kind of weird if I had to show my phone to everyone. But if take their phone and walk around and look at it from their perspective, I guess yeah, it could definitely be interesting and definitely unique. But I don't know because I haven't seen it in action, and having a little bit of a tough time completely visualizing it. But I think overall it would be great because it looks more realistic, it can be just the information that you need in the AR instead of having a lot more information overload with just a screen and all the data, and stuff like that. So, just making it more realistic can make it easier to understand, I guess.

Interviewer: "Alright, that's it."

10.2.2 Participant 2: Fundamental Level

Interviewer: "For a brief introduction, this is for our master thesis where we are doing some preliminary work where we investigate people's preferences and knowledge in statistics. If you

want us to elaborate on any questions, feel free to let us know, and feel free to ask, if you have any questions.

So, the first question I have is: What level would you rate your statistical knowledge to be? We have the fundamental level, intermediate, and advanced level. And if you use any of the topics inside any of the categories you will be considered to be at that level."

Respondent: "So, because I am only at the second semester [Medialogy] I would probably put me at the fundamental level of descriptive and basic statistics."

Interviewer: "So, can you describe your current process for visualizing and analyzing data"

Respondent: "Yeah, uh, we use normal things like histograms and graph bars and such tools to describe for example how many students we have tested and then what their answers has been on a scale from 1-5 or 1-10, for example.

Interviewer: "So, you are using Likert scales?"

Respondent: "Yeah, Likert scales."

Interviewer: "What tools do you use?"

Respondent: "We use Python, Python notebooks, generally to describe the data by making it into diagrams."

Interviewer: "And what type of dataset do you find yourself using most often? Is it pie charts, box plots, heat maps, scatter plots"

Respondent: "For now it has been mostly scatter plots, but we also use histograms quite a bit.

Interviewer: "Okay, do you find it difficult to interpret scatter plots?"

Respondent: "I mean, it is very dependent on the data, I would say. It can, if your data is very split up from all the different kinds you are getting it from, it can be kind of not easily deduced what it actually says. But, it is very dependent on what you are studying and how many you got into the study.

Interviewer: "So, can you please discuss any challenges or difficulties you face when having to visualize data?"

Respondent: "Well, because we are using Python and we are coding how the diagrams should look it can be a bit annoying to adjust how much you should show on it. It always has like a preference on how much it should show, but not necessary the scope you want to measure within, or it gives a slanted view of the data that you don't really want. So, it doesn't really necessarily show the points that you want from the data even though it is within the data."

Interviewer: "So, that can be some kind of visual bias?"

Respondent: "Yeah, exactly."

Interviewer: "Okay, uhm, have you experienced cluttering or occlusion, or anything like that?

Respondent: "Cluttering, and what was the other thing?"

Interviewer: "Occlusion."

Respondent: "Occlusion, can you elaborate on that?"

Interviewer: "It occludes so, if I have my hand here, and I take my hand from here to here, it is occluded, it is behind it, but you can see it again here.

Respondent: "Ah, okay yeah. Uhm, yeah kind of, if you compress the data a lot it can kind of obscure some of the actual content of the diagram. So, it can definitely obscure parts, I would say.

Interviewer: "Do you think interactivity and real-time manipulation would help you in visualizing your data?"

Respondent: "Probably, if there is a tool that easily makes me able to set it up so I can design length of the how much i want to test and the amount of test subjects and amount of answers. If I could easily scale that it would make it easier."

Interviewer: "So, in a bit of another direction. Have you ever used augmented reality before?

Respondent: "I have not."

Interviewer: "Have you tried using Pokémon Go?"

Respondent: "I have tried that, yeah."

Interviewer: "So, you do know kind of what augmented reality is?"

Respondent: "I know what augmented reality is, yeah."

Interviewer: "Okay, so, would you consider using augmented reality for visualizing and manipulating your data?"

Respondent: "I can definitely see how it would be useful, because the human body is intuitive to use, so if you could just look at graphs and then kind of expand them or such with your hands, it would make a lot of sense."

Interviewer: "So, you would like the interactivity to be part of your augmented reality experience for visualizing data?"

Respondent: "Yeah."

Interviewer: "How do you think augmented reality could enhance your current data visualization and your workflow?"

Respondent: "Well, it would make it easier to get rid of these roadblocks of - I need to write specific code to make it be the exact size I want it to be and the right scale. I would think it would

just be that I like pull them apart to make it more spread out, so it would make it more fluent, or make it flow better.

Interviewer: "Do you think it would be more engaging or immersive to use?"

Respondent: "If it has animations to compliment it or some visually, then definitely."

Interviewer: "Okay, so in the context of 3D data; do you think it could be more intuitive to understand data visualizations by moving a mobile camera than a traditional mouse and keyboard?"

Respondent: "Hmm, yeah I would say so. Because a mouse it an abstract tool whereas when you are dealing size and actual move a phone you get a sense of actual size in the real-world, so yeah it would be more intuitive.

Interviewer: "You also get allowed to use the degrees of freedom you have, when you move around the mobile compared to the mouse, where it is very static in terms of having one specific point. Uhm, so if you were to measure the success or effectiveness of an AR data visualization tool, how would you measure the success?"

Respondent: "Of the tool?"

Interviewer: "Yes."

Respondent: "Well, first of all I would always give it further along the process, I would always give it to somebody to see if they find it intuitive to use, first of all. Then I would see, if, how much animation is needed before it gets too distracting. And then also just a lot of how it feels, because if it is an AR thing, you have a lot of flow within moving your phone, so I would want to get rid of all the hiccups where it doesn't follow for a second, or something like that."

Interviewer: "So, you think it should be easier to use, and it should be easy to learn?"

Respondent: "Yes. Because I think people who already know a lot about statistics they can do all these things themselves, however, it is more for somebody who has a hard time knowing what do I use statistics for that would need this application."

Interviewer: "What about the complexity or simplicity when it comes to specific features?"

Respondent: "I think you can either do it as a thing where you get more features as you get different diagrams. So, your knowledge on the subject would increase as you use the tool more and more, and then you could visualize your data in more and more ways, and therefore it has a higher usability."

Interviewer: "How would you prioritize the just mentioned measures from most important to least?"

Respondent: "It is most important that it is easy to use, and that is has flow. Then, having more features come next. The more statistical things I can actually visualize with it the better I think

this would be, as a general tool."

Interviewer: "Lastly, how would it be helpful to use data visualization in augmented reality, or not how to, but would it be helpful to use data visualization in augmented reality as a presentation or explanation medium? So, if you were to present your data work to some study members, do you think it would be more optimal to use an AR medium compared to a 2D medium?"

Respondent: "It depends on how many people you want to show it to. I think, if it's to a few people it can be very easy, because you can always just give the device to the other person, and then they can interact with it themselves. However, the more people you get into it, I think it would be hard to show a lot of people because a lot of it is interacting with the device, and if I have to wait before 25 other people have interacted, then it kind of falls a little flat, I think."

Interviewer: "Alright, that's it."

Respondent: "Yeah?"

10.2.3 Participant 3: Intermediate Level

Interviewer: "So, for a brief introduction this is a master thesis project where we develop an augmented reality application that uses data visualization, and this interview is part of our preliminary work. And we want to investigate people's preferences and knowledge in statistics. If you have any questions feel free to ask.

First and foremost, we want you to kind of rate your statistical knowledge, and we have based on three different categories, one of which is the fundamental level, intermediate level, and advanced level. If you use any of the methodologies in the parenthesis then you will be considered to be at that level."

Respondent: "Good, definitely in the intermediate level."

Interviewer: "Alright, for the next question. Can you please describe your current process for visualizing and analyzing data?"

Respondent: "Yeah. It depends on what you mean about the process because it actually starts from... It is very important when you do this that you have a very good foundation. So, in the research design also think of how to visualize things in the end. So, it also depends on the... which outcome is it that you would like to have. And that depends on which kind of data analysis you go along with. So, it is the same within quantitative data, there are certain criteria for whether you are doing an ANOVA test, t-test, Mann Whitney U test, and so forth. But it is also a matter of which outcome you would like to gain. So, it is as much about having already that from the beginning in the research design in terms of trying to outline and more specifically in terms of which outcome is it you would like. Or what is it you would like to know. And based on that, it is moving on into the process of which analysis you are going along with. So that would be kind of my answer, so it goes maybe a little bit back in the process starting with the foundation."

Interviewer: "What tools or software do you typically use?"

Respondent: "Well, I... In terms of visualizations?"

Interviewer: "Yes."

Respondent: "So, that's mainly Tableau and Atlas.ti, but also Excel."

Interviewer: "Which datasets do you find yourself using most often?"

Respondent: "Most often it is kind of minor datasets in terms of relatively few participants in terms of the statistics. But it can also be quantitative data with interviews. But previous research has been with you know 8000 kind of surveys. No, that was 8000 observations. So, observational data. 2500 participants in the survey and interviews and all that in terms of a more mixed method strategy. So, it can also be combined within both the quantitative and qualitative. But I have experience with both quantitative and qualitative.

Interviewer: "What about the graphs and plots? What do you typically tend to use?"

Respondent: "In Excel it could be the box plot for instance, very traditional kind of frequency, communicative frequencies and studies. And then extracting it in Excel by numbering. So, Excel within box plots. More advanced I would say, used in qualitative research, within for instance Tableau for having visualizations of frequencies of mentioned words in bubbles, for instance.

Interviewer: "Do you have experience in using scatter plots?"

Respondent: "Yeah."

Interviewer: "How so?"

Respondent: "It is also mainly used within Excel."

Interviewer: "Can you discuss any challenges or difficulties you face when having to visualize data?"

Respondent: "Yes, there's a lot of things, because I think it is a core of what you would like to communicate. Because there is also kind of there is the devil. Do you actually visualize the data in terms of what you would like to communicate? Uhm, so I think there are something in there that is quite interesting. So, in terms of the validity and the reliability, so does it actually represent in terms of what it should represent? Or what is the core elements in it? So, and also sometimes it can be a matter of, so if you have bubbles and stuff like that, what are the relations to each other and what does it actually say? So, you can have a huge large bubble, right? But, what is that? Is it 55 percent or is it 55 in numbering. I think it is very important to be specific in terms of the labeling, but also in terms of what it is against so to speak. So, is it based on 10, you have the big 10, 10 participants or 10, I don't know whatever numbering as the large bubble, and then a medium bubble, 5, and then a small bubble, 1, right?. So, what are the, so to speak, what are the n, the big population, and what are the little n, the sample size? So, I think there are often some

confusion in what we are talking about. And also in terms of the labeling, they are not specific enough."

Interviewer: "How would you typically address these challenges?"

Respondent: "By being specific. But also being very cautious about the difference between the population and causality, because it is not necessarily that there is an effect-relationship, right? Due, there is a correlation. So, there could be a correlation, but it does not mean necessarily that there is a real effect in terms of the whole causation.

Interviewer: "Do you think that interactivity and real-time manipulation would help in data visualization?

Respondent: "Yes."

Interviewer: "Do you think it could help to highlight or change scale of axis', colors, etc?"

Respondent: "Yes, it could be. It would make sense. So, we are also in a project about visualizing in VR - math in terms of vectors. As vectors are in 3D then it makes sense, maybe to have it in VR where you can actually move around and see things in 3D so there is a one-to-one match between the 3D elements and what it is you would like to learn. So, that is kind of an interesting project."

Interviewer: "A bit in a different direction. Have you ever used augmented reality before?"

Respondent: "Yes."

Interviewer: "Would you consider using augmented reality for visualizing and manipulating your data?"

Respondent: "It depends. In terms of the foundation, in terms of who are the users and which context is it that you have, and also in terms of learning objectives. So, what are the success criteria, so what are you aiming at? So, I would rather choose kind of the knowledge before the method. What is it that you would like to gain, and then decide in terms of which technology you are going. Is it actually the most fruitful of using AR, or are there any other, better, kind of technologies that could be more beneficial in terms of what would like to be."

Interviewer: "How do you think augmented reality could enhance your current data visualization workflow?"

Respondent: "I think, maybe yeah, maybe it could be more immersive and represent kind of things in other ways that you could actually maybe having it... also as a training tool. Something step by step, do this, do that. It could be that you could go back including different things - now you do this, now you do that. It could be as a training tool might be possible. But there are also some disadvantages."

Interviewer: "So, in the context of 3D data; Do you think it could be more intuitive to understand the data visualization by moving around a mobile camera, compared to a traditional mouse and

keyboard?"

Respondent: "Yes. But then again it depends on what you are aiming at. What is it you would like to communicate? So, not necessarily. It depends on the user and the context."

Interviewer: "So, if you were to measure the success or effectiveness of an AR application for your use, how would you measure the different aspects? What becomes most important for you when you have to use a new application?"

Respondent: "I would look at... There would be different criteria I think. So one could be looking at the users. How well the users perform in certain criteria they are performing in different ways, right. So, that could be one thing. But I also think it is about the user acceptance actually. So, in terms of whether the user find it useful and whether the user find it worthwhile using time spend on this, is it actually worth using it, and also, can it be recommended to others - that could also be a criteria. Is it so good that you can also recommend it to other users? So, all this within the technology acceptance model from Davis and others."

Interviewer: "What about the usability?"

Respondent: "Yeah, of course. But that is also an included part of it. For having it useful it should work, right? Not to many bugs, and not to many figuring things and all that. It should just work."

Interviewer: "So, if you should prioritize from the most important aspect to the least, how would you prioritize the different aspects?"

Respondent: "What are the scales?"

Interviewer: "It is up to you."

Respondent: "Okay, uhm. User and context. Most important. Pilot testing. I think iterations are also important, when you develop things to improve. But, I think the most important thing is to include the users and the context from the very start, and then also to have some success criteria as well as to have it as a real project, as a real societal impact. I mean, why develop something that is not useful?"

Interviewer: "What if you think of it more in a way that you are to use this application, how do you measure the success, how do you find it satisfactory to use?

Respondent: "Yeah, that could be different things. It could be looking into the engagement, it could be within the user engagement scale, for instance. Whether they have the attention, the focused attention, it could be within the usability, it could be within whether they find it worthwhile, but there could also be included some knowledge test. Aesthetics and visualization is also important of course. There could also be something within the acceptance of the technology. Yeah, there could be other things.

Interviewer: "And for my last question, do you think it would be help to use data visualization

in augmented reality as a presentation medium?"

Respondent: "Yes, but it depends again on the users and the context and what aim you are having."

Interviewer: "Can you provide an example of a case where you might see it being useful?"

Respondent: "As I said, as a training a specific element, maybe for specific training elements for statistics it might work, including some visualization within statistics could work. Guidance, training, help for the users to provide a here-and-now kind of instant training - do this, do that - that could be a thing, and you know you could also include it with a video if it is too difficult to manipulate or to have it as an included design. You can also make a video. There are many different things that you can do."

Interviewer: "Alright, that's it. Thank you very much."

10.2.4 Participant 4: Advanced Level

Interviewer: "So, for a brief introduction: We are a master thesis group, and we want to develop an application for data visualization, and this interview is part of our preliminary work, where we want to investigate people's preferences and knowledge in statistics. Feel free to ask any questions, and let us know if you want us to elaborate on something.

So, the first step is to rate your statistical knowledge. So, based on our three categories from: fundamental, intermediate, and advanced - based on the criterion of these categories, how would you rate your knowledge?"

Respondent: *Circles* the advanced option.

Interviewer: "So, can you please describe your current process for visualizing and analyzing data?"

Respondent: "I typically obtain some of the cloud services coming out from our university, and then I use, basically, for the tasks which are there, SciKit learn for simple problems. And whatever the problem, I make a simple Jupyter Notebook on the server and put my data there. Usually, using s-copy but also sometimes even drag-and-drop. This data, the latest data I used was made up data on some students' grades. So, after I have data there, I used Scikit learn to read the data and simply use matplot lib type of plotting functions to provide me some entry for the tools. If I'm dealing with very simple data and I want to try out as much as possible methods on visualizing the data, then I could employ sometimes higher level packages, like the PyCaret. It runs through a lot of tests, a lot of predetermined steps to make exploratory data analysis. Then with one comment I got lots of futures and a little bit of principle component analysis-type of things to reduce the dimensionality, and it also gives me a 3D plot of several aspects of the data using T-SNE or something. But usually, after analysis what I typically do is to turn them into simple two-dimensional graphs that I can try to understand and get something from the data. That is my

general workflow. There is some specific task depending on the data type or the problem. But, usually that is my starting point. Either very simple matplot lib plots, or using AutoML to get a lot of futures without too much programming, so I can get started."

Interviewer: "So, what kind of datasets do you find yourself using most often? Is scatter plots, pie charts, box plots?"

Respondent: "Usually, scatter plots are the go to things, when I look at the dimensional analysis. Pie charts occasionally, when I want to give out some distribution in very simple terms. But usually, scatter plots are my first go-to."

Interviewer: "Can you discuss any challenges or difficulties you face when having to visualize your date?"

Respondent: "Of course, uhm. I mean, the most challenging is of course with the multivariate data that there are a lot of aspects of the data that I need to look, and I want to understand each aspects are probably the most important once would explain me the majority of variance that I train machine learning to. So there, we have developed some very simple dimensional analysis, but if it is more complicated than that then I use PyCaret to automatically look at the importance of the data. So the most important thing is multivariate data. Different data types are of course also challenging. If I have a labeled data - how to convert that to scatter plot because there are different types of data, sometimes I have to do the data conversion, and that's another challenge."

Interviewer: "How do you typically address these challenges you just stated? How do you try to fix it?"

Respondent: "If it is easy, if I can do simple approaches like ?one hunt? In coding or something like that to provide for each class type another either float or integer dimension then I use that. But, if it's more complicated than that, PyCaret gives me a way to do that. Again, higher level package to simply that data processing."

Interviewer: "Do you think that interactivity and real-time manipulation would help you in data visualization?"

Respondent: "Oh yes. Of course in many cases it would. One thing I can do there is... include faster API type of approaches which gives me a way to at least parameterize data and show different things, but if it is more complicated than that I in very seldom cases, in several projects, I use a bit interactive visualization like Voila-type of approaches but it is very seldom in what I do. Because usually I get by after a little bit of looking at the data visualization for exploratory data analysis and after that I typically go to number crunching rather than more investigations."

Interviewer: "Okay, so in a bit of another direction. Have you ever used augmented reality before?"

Respondent: "Yeah."

Interviewer: "Did you enjoy using augmented reality?"

Respondent: "Uhm, in some applications, yes. In some applications it was a superficial element that it didn't bring to much value. But in some, definitely."

Interviewer: "Would you consider using augmented reality for visualizing and manipulating your data?

Respondent: "Never tried. Never tried for pure data things. I always tried that with the reference to spatial coordinates, like maps and stuff like this. And encoring 3D objects to the floor. But never for abstract data visualization. Probably, yeah, there might be interesting cases that could benefit. But I haven't tried myself."

Interviewer: "Okay, how do you think augmented reality could enhance your current data visualization workflow?"

Respondent: "Let's say, if I would work with very sophisticated graph data, for example, which also has this spatial relationship, of course I would like to rotate, manipulate the data in 3D. Any kind of molecule protein structure, any kind of located graph such as, maybe edge or mobile computing devices which always form a graph. It might be very useful to know this 3D representation of this structured graph for understanding what is the temporal relationship in which state they are active, or in protein folding like of applications. Their spatial should be important as well. In that I would like to manipulate data - zoom in, zoom out 3D rotations. I think if I would deal with that kind of problems, I would strongly benefit to do that in augmented reality."

Interviewer: "Do you think it would be more engaging and immersive?"

Respondent: "Yes."

Interviewer: "So, in the context of 3D data; Do you think it could be more intuitive and easier to understand by moving around a mobile camera than your traditional mouse and keyboard, in order to rotate and zoom?"

Respondent: "Possibly yes."

Interviewer: "Would you see a benefit from doing it?"

Respondent: "Now I'm really thinking, so if I think about a simple graph which has a root, then there's a branching of almost in binary form. Then I don't see too much of a benefit to have an external camera because trajectory is very clear and I could get as good results with mouse and keyboard compared to the external camera. But if that does not hold like in the protein form, it's not a binary thing, it can be any type of branching, any difficult 3D transformation structure, then I would like to have something that I can manipulate easily with my hands, like the camera angle, compared just very simple mouse and keyboard interactions. So there might be cases. If I work with more complex data and maybe even also the new virtual reality objects, or something. Very unstructured, very spatial data, then I would definitely benefit from direct manipulation."

Interviewer: "So, if you were to measure the success or effectiveness of an AR data visualization tool, how would you measure it?"

Respondent: "Probably, I would definitely compare it to a baseline 2D screenbase interaction. And I would really like it to go much beyond that. One way of doing that would be, if i'm dealing again with a protein structure, maybe I can ask to locate a particular molecular structure, substructure, in the whole 3 dimensional things, and compare augmented reality versus 2D representation with mouse and look at both the time it took to find out or give a simple task to manipulate it in a certain way, and I look at the time difference between two different modalities. And I would also, probably, ask the user qualitative parts about how did they consider the task hardness. How they considered the difficulty. Sometimes cognitive, you've already mentioned in your report, this kind of thing like the NASA cognitive load survey. But I would like to get qualitative understanding about how they find the manipulation in both modalities.

Interviewer: "So, if you were to measure the tool specifically, would you kind of determine on the ease of use and how easy it is to learn. What would be, like, personally, what would be important to you, if you were to use a new tool?"

Respondent: "I think the first thing will be the direct issues related to the tasks that I have at hands. Some high level task, and how much time I could use to complete it in augmented reality or in any different modality. If the tasks is taken much harder in augmented reality compared to a screenbase thing, then probably I wouldn't go any further to ask more questions. How did it feel and everything would be secondary to the first objective. But if the things are comparable then I would go ask the questions that you mentioned. How easy it was to learn? After learning did I enjoy it? Can I complete similar tasks as well? Or slight deviations of the tasks easily. I think I would be very interested to know these kind of things. After I get the basis right. Like, substantial change between the modalities for me to complete the task."

Interviewer: "So, if you were to prioritize these success criteria, how would you measure it from the most important to the least?"

Respondent: "I think I would definitely go for... To give a task and look at the task completion time in different modalities, that would be my first thing. And then, if the tasks are at least comparable but also in favorite of the augmented reality modality, then I would go to a little bit secondary more qualitative aspect of enjoyment both particular modalities and that. But I think, still the first one will be task completion time in different modalities.

Interviewer: "And for the last question, would it be helpful to use data visualization in augmented reality as a presentation or explanation medium? So, if you were to present your data to others, do you think it would be better if you used augmented reality?"

Respondent: "Yes, I thinking about problems all the time. So, if there is going to be hard and really spatially important that I said non-binary harder representation then augmented reality would

surely benefit. Then I could also think about very peculiar databases... No but the incensory information is shown to organize in a very, very strange banner. What we call ?Pawn Carry Maps? So... In the sense think about a circle. In the center it is sparse, and when you go to the edges it's becoming more and more dense. And actually the whole horizon of infinities map onto a circle. So, when you go ultimate circle there's an infinite dimension there. Physically on a pen and paper, in 2D this doesn't make any sense. It only makes sense if you can make interactive visualization based on that. That you need to navigate in that space. And there, I don't see, maybe virtual reality could work as well but augmented reality to put that on to the existing environment. I think that is the way to go. So, to your question, I thought a bit aloud, but there is some data if I would work with them, I would strongly think augmented reality data visualization is the first solution to go."

Interviewer: "Alright, that's it."

Respondent: "Nice."

10.2.5 Participant 5: Advanced Level

Interviewer: "As an introduction, uhm this is for our master's thesis where we develop an application that uses augmented reality for data visualization and this interview will be part of our preliminary work where we investigate people's preferences and knowledge in statistics. So, if you have any questions, just feel free to ask. Or if you want me to elaborate on any of the questions, just ask.

So my first question is what would you rate your statistical knowledge to be? Is it the fundamental level, intermediate level, or advanced level? And if you use any of the methodologies within any of the levels, you will be considered to be in that level."

Respondent: "Ok, hmm. I'm a little bit more than intermediate, but I don't know machine learning and I'm just learning Bayesian now. So, I know Time Series. I have studied that, I took a course in datacamp. I also took a course in Bayesian statistics, and I'm really into learning it right now, so that's kind of my current status. And machine learning I really haven't learned, so apart from that, i know all the rest."

Interviewer: "Okay, that's fine."

Respondent: "So it's a little bit like between intermediate and advanced I guess, or advance maybe?"

Interviewer: "Can you please describe your current process for visualizing and analyzing data?"

Respondent: "Well, it depends on the data and also depends on the purpose. Like, I usually use R, and I have like a set of packages and a set of procedures that I follow. If I need to analyze certain type of data or conduct certain type of analysis, I will just go to these procedures I have and check like, okay what do I have to do? Like, I'm very systematic so I just check what i have to do and

kind of run the analysis I have to do based on what you will call in your education as pipeline. Yeah, so that's pretty much what I do. But that if I have a load of data. If I don't have much data, and I need to do something very quickly then I just go to JASP. It analyzes thing very fast and it's very easy. So if I just need to, ok… How, is there any correlation here? I just go and analyze if it's any correlation. If I'm working really for a project, especially if I'm working for a open science framework I really produce, I use R to produce these. I produce something different for producing reports because for the upper science framework you have to be very transparent. So anytime you make analysis, you produce a report, and for that I use a markdown. So they are markdown. I use it for producing reports which are also like you can tie them to your database and then you can make your resource reproducible. So I use R when I'm working in a very, very and important projects. If it's just for me to check something, I just use JASP. About the graphics for R, I use the normal one, that is the GGPlot2 for data. And then depending if I'm going to do like network analysis, there is one that is very specific for network analysis. And yeah, it depends on that. And for JASP, it auto generates the graphics. So yeah, it's pretty much it, but yeah, GGplot mostly if I'm using R."

Interviewer: "Alright. Which datasets do you find yourself using most often?"

Respondent: "What do you mean?"

Interviewer: "Is it histograms, box plots, scatter plots, heat maps? Kind of what type of graph or plot do you use most often?

Respondent: "Well, you know, it always depends on what type of data, but I think histograms and like violin plots and all this like whisker plots, I forgot the name. Those are the ones I use the most because it allows me to see a little bit more on the data. I only use like very specific types of data in… A scatter plots as well if I'm looking for example too, uhm, for correlations between numeric data. But yeah, because that's what you use. If I'm doing obviously network analysis, I use network, or if I'm using the… It depends, it really depends, but the ones that I did data I normally use, the type of nice like normally use are like linear regressions and correlations. So, there's normally boxplots, violin plots and scatter plots. But I also I I took a course back during my PhD in data visualization and they also suggest things like don't use pie charts and stuff like that because they can get really deceiving and they tend to deceive the mind of the reader."

Interviewer: "Yeah, exactly. And this kind of lead me to the next question. Can you discuss any challenges or difficulties you face when you have to visualize data?"

Respondent: "I think the most problematic one that gets, I don't know, like when you are working with data analysis. One of the things that is problematic is a scale, because you can actually deceive people really easy with scale. If you have like a very weird scale, you can show for example that some two things, let's say a comparison in a box plot, and two things that are not so apart from each other are very apart from each other. Just if you change the scale that's actually something i used to teach when I was teaching statistics back in Sweden. And like how you can

deceive people, but just to prevent people from to deceive people with graphs, because people just go for the graphs and they don't really look at the numbers, and many times you see maybe that's not so much done in academia, although I have seen in academia as well but, it's more like in in scientific journalism when they show 2 graphs and they show them like graphically it looks like the separation is super big but statistically is like 2 decimals or something like that. So the scale is really important and it's always very challenging, like OK this scale, is it really like reporting like is visually depicting what the mathematics are saying or am I kind of cheating the viewer, you know? So I think that's one of the biggest challenges when you are visualizing something. Another challenge is to make it clear, particularly when you have like a lot of data and you want to convey a lot of data with one graph that happens a lot when you are publishing in journals because you want to save space and you don't want to have like 5 different graphics that can be conveyed in maybe one. So you try to add information in one graphic. So should I put for example this scatter plots and make also the you know the the dots make them bubbles and then the bubbles give them color and give them size according to different things and then I can convey more data in just one graph. I think that's in art and it's very difficult to get there, but if you have like enough time and enough imagination, you can do some very, very good synthetic graphics for a lot of data. But in general, I think we go for the most simple because it's kind of the safest, yeah.

Interviewer: "So, kind of in a different direction. Do you think that interactivity and real-time manipulation would help you in data visualization?

Respondent: Yeah, interactive graphics are very, very nice, but of course you use them more for like exploratory analysis. You don't use them for in journals or publications or anything like that. I think interactive graphics can be very cool for learning statistics. That is something that I will totally love. When I have been teaching statistics it's something i will totally love to have also because normally I teach statistics to people that don't know mathematics, so making it more like more playful like you can play with the data. I think that would be great. The other thing is that what I said is more for exploratory analysis, but it also depends on your take, because I have also seen people doing exploratory analysis just to like checking which data correlates with what. For example, just so they can fish questions, which means like they are doing some kind of harking with that. But in general, if you really have like an open research question, like if he's there, any correlation between any of these variables and these other variables is really is really good. And I have tried, I think excel for example has a function like that. That you can just like move data from here to there and then it changes the graphic and I think R also, when you are real time you can also do like graphics that that kind of have an animation. But apart from that, I have not really played much with that. As I said, it will have been really, really cool to have it for classes. But, but I haven't played much with it because it's normally not what you publish and when I do it, it's mostly for publications, but I think it will be very useful for teaching and for exploratory data analysis, yeah."

Interviewer: "Do you think that animations would help the kind of engagement and the immersion of data visualization?"

Respondent: Uh, yes, especially if you are learning statistics, but also if you are for example, in a different environment, let's say a conference, because of course publication is is one thing is support on a paper, but if you are in a conference and you can show how your graph for example, let's say time series, how something is evolving over time and instead of showing like the how it started and how it ends, you show the whole process like changing kind of unreal time on screen that has way more impact on the public and also it gives way more understanding of how actually the process is. So, those kind of things I think are very important, yeah."

Interviewer: "Okay, in a bit of a different direction here in order to touch upon our different theme for this project, have you ever used augmented reality technology before?

Respondent: "It depends how you are asking. I have used virtual reality to check on your projects. So when you have virtual reality project, I use virtual reality. I haven't programmed or done anything virtual reality, at least not yet. And I don't use it for recreational purposes because I get a motion sickness very much from it. So I never really bought a set for myself.

Interviewer: "Did I say virtual reality?"

Respondent: "Yes, I think so."

Interviewer: "I meant augmented reality."

Respondent: "Augmented reality, uhm, I think the only augmented reality I have used is Pokémon Go. I haven't created anything in augmented reality myself. But I think Pokémon Go is kind of the only augmented reality I have used. I have played a little bit with it with some things like I remember my my 3DS also had an augmented reality game. Actually a set of games that I also played but apart from that it is not like I go to augmented reality often."

Interviewer: "Okay, but you know what augmented reality is?"

Respondent: "Yes."

Interviewer: "Would you consider using augmented reality for visualizing and manipulating data?"

Respondent: "Hmm, I think it depends on the setting. Also, because augmented reality normally requires you to kind of, at least the type of technology I have used, is like bind to the telephone. So you have to take your telephone. I think it will be very difficult for you to kind of hold the telephone with one hand and manipulate the data with the other. If there are other ways like glasses or something that allows me to have like free hands and see the data in front of me like in the heavy rain game, if there was something like that, I will totally go for it. Like seriously. But the thing is that it should be like augmented reality that goes directly to my eyes, not something that I have to hold with my hands."

Interviewer: "How do you think augmented reality technology could enhance your data visualization workflow?"

Respondent: "I really don't know to be honest. Because I'm also used to work on the screen and you know you make the script you run and you get your graphs. So I have really no idea if I... I'm trying to put myself in that position, but if I'm manipulating, how can I manipulate my data like the data has to be also like visual right? So if I put like all my, let's set a bowl that represents my data and combine it with another bowl that is another type of data with my hands and then I just say that may like a scatter plot of those two types of data. They said they are variables and then I can see that. I think that will actually not add anything. I think it will make things actually more convoluted than they actually are, but if I have to explore 3 dimensional data or data with more dimensions than two, which is something that we don't do usually, like if you are making, you know, let's say a heat map or something like that, you want to represent it more in three dimensions. Then maybe will be very cool, because then you can rotate it along the axis. And see, actually where the big points are instead of just looking, you know, at the colors. So I think maybe for data that is more than two dimensional, so three and maybe more dimensions, I think that will be very useful. Like now, I obviously I haven't done it so. So I'm just imagining myself in which situations it will be useful. I think that will be useful, but if it's like for normal correlation data or box plots or anything like that, I don't think it will add anything."

Interviewer: "You kind of answered my next question already, but just to follow the procedure. Do you think in the context of 3D data, that it would be more intuitive to understand the data by moving around a mobile camera than a traditional mouse and keyboard to rotate and zoom and so on?"

Respondent: "Yeah, I think it will be a little bit easier, yes. Maybe a little bit more intuitive. But again, when we talk about, when I talk about 3D data, I don't mean 3D graphics. I just mean 3 variables data, and I have to be specific on that because 3 dimensionality that's also something I learned in the course I took of data visualization 3 dimensionality like making a box plot in 3D, like that you can do for example in Excel. But your representation in 3D can be very confusing for the brain and it's also a way to mislead the reader of what kind of data you have produced. So normally 3D graphics are kind of you should never use them for papers and things like that. I mean with 3D data is that it is 3 variables or more data. And I really think, like if you manipulate it with your hands, it will be more intuitive than with mouse and keyboard. Because it's kind of convoluted with mouse and keyboard. We have all been there. We have all used Unity and Blender.

Interviewer: "Thank you for clarifying. If you were to measure the success or effectiveness of an augmented reality data visualization tool, what would your measurements be? Would it be that it is easy to use, easy to learn, or are there some other categories that you think are more important?

Respondent: "I Think that the first thing will be that it is easy to learn and easy to use. For me,

the comparison should be with a tool that is already existent. If I find more convoluted to create and explore a graph in augmented reality than I'm used to, for example in R or in in any of the of the software I use, then of course, why should I use it if it's more difficult, it's maybe another learning curve. I'm not going to say more or less because I don't know, but maybe it's another learning curve. Another thing to do just for having results that I can have easier in the software I use. So I will say that usability is really the key point. How easy it is to use and how easy it is to manipulate, because that's kind of also the added value if I can manipulate it easier with the hand and maybe show it to other people in an easier way that maybe sharing a video. Because normally you have to share things like videos or GIFS to people if you want to show them something like rotating, else you will have to, you know, share the file or something. The original file, and that's that's more convoluted. That's not very good for communication, but if you can do it kind of in real-time like that. So the type of communication as well I think is is an added value, but me as a scientist, like if I'm focusing communication is a very good thing to have, if I'm focusing as a scientist, I think it should really be easier to use than any of the tools I use and it should give me more things like for example if I can see three dimensional data in a way that I cannot with my current software. That's very like that's another value, right? Even if it takes maybe a couple of extra steps, so that's what I will say, like easy to use and that adds something extra to what I already use. And I'm not going to say so much like easy to learn. Of course it will be very good if it's easy to learn. Especially, like if you're a professor or or a scientist, you don't have much time to learn a new tool, but I will say it is in the extra things. What does it give me that my current tool cannot give me? That's what I will focus on."

Interviewer: "Okay, so if you were to prioritize what you just mentioned from most important to least, how would you rank it?"

Respondent: "First of all, usability, so it has to create or do something that is as easy. Ease of use. That's the first one, so it should create something that is at least as easy as the software I already use. Uhm, else it is like learning a whole new thing, and that will be kind of annoying. And second, it should allow me to like it should have extra. It should give extra things like for example that I can use it at a conference to show people my data and that they can interact with the data that would be actually awesome or that I can use it to record a video with it and show it at a conference like put it as a GIF or video in a PowerPoint presentation that will be also awesome. And the third one is that is has like a short learning curve but I kind of prefer a smooth learning curve than a short one because normally short means also steep, but that it is easy to learn that will be the third priority."

Interviewer: "Do you think it would be helpful to use data visualization in augmented reality as a presentation or explanation medium?"

Respondent: "Oh, totally, totally, I think it's… if I particularly me, because I'm interested also in scientific communication, if I will use it for something will be for that. And also because you cannot really use it for papers like I cannot put an augmented reality in papers. Maybe for your

project or something you plan to make it like that people can access it via QR code or something, and maybe you can do that and then other people can see it somewhere and kind of visualize it like you put it in your paper as a QR code and then people can visualize your data there with their phones. That will actually be pretty cool, because then you they will have the data and kind of maybe play with the data or play with the graph and you know, augmented reality there with your paper in front. That would actually be very cool! So instead of putting graphs, you put QR codes and then the graphs just come with your telephone. Actually, that's a very, very cool idea. So yeah, apart from that, I wouldn't say that. Like you could use it in a different way like because I don't know how to say this. in a classic paper, let's say that we go away from the idea of that you are codes in a classic paper. You just put your things there, like it's printed. The things don't move. Even I that work with memes and stuff, it's always, you have to put a GIF but the GIF is static. It is not moving because it is a PDF and that's actually takes away from the meme in itself. But if you could put it like if you could put the graph in a in such a way that people can see it anywhere, like in the QR code example or in the in the in the example of presenting it like, go to a presentation and then allow the public, for example, to interact with your data via this augmented reality. That will also be like very, very cool for me. A very good way to communicate your results, I think."

Interviewer: "Alright. That was actually it for the interview. Thank you for the answers.

Respondent: "No, thank you for inviting me."

10.3 Appendix C

Experimental Design Protocol:

1. (Researcher): Welcome. Today you will test our augmented reality (AR) data visualization application. You will have to visualize some data using the provided iPad. We will ask you to complete different tasks while using the application. After completion of all tasks, five in total, you will be asked to fill in a 26-item questionnaire about your experience. Feel free to ask, as you go along if you have any questions.

2. (Researcher and User): Equip EDA & Equip BVP & **(Researcher):** Draw grid (iPad).

3. (Researcher:) Measure baseline for EDA and BVP.

4. (Researcher:) Give participant the iPad (if experimental group).

5.1. (Researcher:) Experimental group: With the iPad in hand, you are able to move around as much as you like to, to visualize the data. Each time you complete a task, you will be asked to fill in a 7-item questionnaire. You can adjust the size of objects in the settings menu. You can also highlight points for additional information.

5.2. (Researcher:) Control group: You are able to zoom in, zoom out, and rotate the data visualization as you like. Each time you complete a task, you will be asked to fill in a 7-item questionnaire.

You can adjust the size of objects in the settings menu. You can also highlight points for additional information.

6. (User) Load Task 1 and solve the task. Press Finish Task after completion.*

Task 1: Identify correlation between Quality and Alcohol (Wine Dataset).
Correct Task Answer: Positive Correlation.
Difficulty of Task: Medium.

7. (User): Answer Task 1 and fill in the NASA-TLX Questionnaire.*

8. (User): Load Task 2 and solve the task. Press Finish Task after completion.*

Task 2: Identify a correlation between Living Area and Price (Housing Dataset).
Correct Task Answer: Positive Correlation.
Difficulty of Task: Medium.

9. (User): Answer Task 2 and fill in the NASA-TLX Questionnaire.*

10. (User): Load Task 3 and solve the task. Press Finish Task after completion.*

Task 3: The Dark Knight was the highest grossing movie in 2008. Find and highlight that data point, and figure out what the exact release date was (Movie Dataset).
Correct Task Answer: 18-07-2008.
Difficulty of Task: Medium.

11. (User): Answer Task 3 and fill in the NASA-TLX Questionnaire.*

12. (User): Load Task 4 and solve the task. Press Finish Task after completion.*

Task 4: Identify a correlation between Petal Length, Sepal Length, and Petal Width (Iris Dataset).
Correct Task Answer: Positive Correlation.
Difficulty of Task: Medium.

13. (User): Answer Task 4 and fill in the NASA-TLX Questionnaire.*

14. (User): Load Task 5 and solve the task. Press Finish Task after completion.*

Task 5: Predict what BMI Classification a person would have if they are 170 cm tall and weigh 100 kg (Height & Weight).
Correct Task Answer: 4.
Difficulty of Task: Hard.

15. (User): Answer Task 5 and fill in the NASA-TLX Questionnaire.*

16. (Researcher and User): Unequip EDA and BVP Measurements.

17. (Researcher): Save Physiological Data as File.

18. (User): Answer UEQ Questionnaire.

19. (Researcher): Switch EDA Sensors and submit observation notes.

* is used to illustrate where the researcher must take observational notes, and note a user's task answer.

(Researcher) is used to illustrate the parts in the procedure where the researcher was involved.

(User) is used to illustrate the parts in the procedure where the user was involved.

10.4 Appendix D

Iteration 1 usability test results
See the supplementary Appendix.zip file attached to this report.
Folder and File name: \Questionnaire Responses\First Usability Test\First Usability Test.csv

10.5 Appendix E

Iteration 2 usability test results
See the supplementary Appendix.zip file attached to this report.
Folder and File name: \Questionnaire Responses\Second Usability Test\Second Usability Test.csv

10.6 Appendix F

Iteration 3 usability test results
See the supplementary Appendix.zip file attached to this report.
Folder and File name: \Questionnaire Responses\Third Usability Test\Third Usability Test.csv

10.7 Appendix G

Experimental NASA-TLX test results
See the supplementary Appendix.zip file attached to this report.
Folder and File name: \Questionnaire Responses\Final Test\NASA Task Load Index (TLX)\NASA-TLX - 5x.csv

10.8 Appendix H

Experimental UEQ test results
See the supplementary Appendix.zip file attached to this report.
Folder and File name: \Questionnaire Responses\Final Test\User Experience Design\UEQ.csv

10.9 Appendix I

Data analysis scripts
See the supplementary Appendix.zip file attached to this report.
folder: \Data Analysis Scripts

10.10 Appendix J

Physiological test results

See the supplementary Appendix.zip file attached to this report.
Folders: \Physiological Measures & \Physiological Measures\LSL_Data

10.11 Appendix K

Unity Project

See the supplementary Appendix.zip file attached to this report.
Compressed Unity Project: Med10-Project-Data-AR-V2.zip