The Rate of Rates: A Study in User Rating Percentage and Satisfaction Based on Design Choices

Hanna Bayerlein and Fredrik Lütz

DEPARTMENT OF DESIGN SCIENCES
FACULTY OF ENGINEERING LTH | LUND UNIVERSITY
2020

MASTER THESIS

The Rate of Rates: A Study in User Rating Percentage and Satisfaction Based on Design Choices

Hanna Bayerlein, Fredrik Lütz

Department of Design Sciences
Faculty of Engineering LTH, Lund University

Supervisor: Joakim Eriksson

Examiner: Mattias Wallergård

February 27, 2020

Published by

Department of Design Sciences

Faculty of Engineering LTH, Lund University

P.O. Box 118, SE-221 00 Lund, Sweden

Abstract

User experience and customer satisfaction are becoming more and more important in the development of digital services. Users are accustomed to easily understandable workflows and fast response times, which is then easy to demand from these services as the range of services and thus the competition between them grows every day. Therefore, in order to both attract new customers and retain the current customers, Telavox are investing in user feedback implementation to keep track of how their users experience their service. This report focuses on analysing the best ways companies can collect customer feedback without being intrusive and time-consuming, while at the same time yield a high response rate.

To answer this, a digital test application was developed where the most common user feedback rating scales were presented and evaluated based on three different categories; Grading Quality, User Satisfaction and Ease of Use. The test was followed by two questions where the test person would answer which of the rating scales they would like to use again and if there were areas of improvement for that scale.

The result of this test was compiled and out of 21 original rating scales, the top candidates were once again iterated into three new prototypes, where the design was improved and then a final user test was carried out in Telavox's own digital application Flow. Here, the test subject was asked to rank the three rating scales and to list advantages and disadvantages for the different systems.

After the final user test the results were compiled, and the winning rating scale underwent another design iteration, which was then implemented in Telavox's product Flow where it is to be used to collect the user experience regarding the quality of the call.

Keywords: Customer satisfaction, User rating, Rating systems, Call Quality, User Experience

Sammanfattning

Användarupplevelse och kundens nöjdhet vid användning av mjukvara blir en allt viktigare del av utvecklingen av digitala tjänster. Vi som användare är vana vid lättförståeliga flöden och snabba responstider, ett krav som ställs allt mer för digitala tjänster, då utbudet och konkurrensen mellan dessa växer varje dag. För att både attrahera nya kunder och behålla de nuvarande, satsar därför Telavox på implementering av kundåterkoppling, för att hålla koll på hur deras användare upplever deras tjänst. Denna rapport fokuserar på att analysera de bästa sätt företag kan samla in kundåterkoppling utan att vara påträngande och tidskrävande, samtidigt som en hög svarsfrekvens kan utvinnas.

För att få svar på detta utvecklades en digital testplatform där de vanligaste betygsskalorna för kundåterkoppling presenterades och utvärderades baserat på tre olika kategorier; Hur väl man kan uttrycka sig själv, Attraktivitet och Enkelhet. Testet följdes av två frågor där testpersonen skulle svara på vilken av betygsskalorna de helst skulle vilja använda igen och om det fanns förbättringsområden för den skalan.

Resultatet av detta test sammanställdes och av 21 ursprungliga betygsskalor gick toppkandidaterna vidare till ännu en iteration till tre nya prototyper, där designen förbättrades för att sedan kunna genomföra det sista användartestet i Telavoxs egna digitala applikation Flow. Här bads testpersonen rangordna de tre betygsskolorna samt ge för- och nackdelar mellan de olika systemen.

Efter resultatet av det sista användartestet sammanställts undergick den vinnande betygsskalan ytterligare en designiteration för att sedan implementeras i Telavoxs produkt Flow där den ska användas för att samla in den upplevda samtalskvalitén.

Nyckelord: Kundnöjdhet, Användarbetyg, Betygssystem, Samtalskvalitet, Kundåterkoppling

Acknowledgement

We would like to thank our supervisor at LTH, Joakim Eriksson, for all his support and ideas during this master's thesis.

A big thank to Telavox, especially our supervisor, Simon Wallström, and Henrik Thorvinger, for giving us your time and help to answer all our questions and discuss our ideas.

Also from Telavox, Johan Westerlund, Sofie Eliasson and Martin Larsson, thank you for your technical expertise and helping personalities.

Finally, we would want to thank all those who have taken their time to help us by giving us feedback or have taken part of our usability testing.

Division of Work

The general division from the start was Fredrik in charge of the front-end development and design and Hanna in charge of the back-end development. The report was not divided beforehand instead both of us would write the chapters we felt were most needed at that phase of the work. But before a solution was reached for each new task, both code wise, design wise and report wise, we first discussed the problem, trying to see it from different angles and perspectives, so that we both always knew which direction our work was headed. The workload between us was also always re-evaluated on Mondays so that the planned deadlines would be met by prioritising the right tasks for that week.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

User Experience (UX) and customer satisfaction play two important roles in today's battle of users. User experience can be defined as described in ISO FDIS 9241-210 (Bevan 2009): ”A person's perceptions and responses that result from the use and/or anticipated use of a product, system or service.". The chance of the user continuing to custom the specific product increases if the experience is good enough to be satisfactory for the user. Therefore, many of the IT solutions today collect feedback from their users, to help prioritising the content uniquely for each user while at the same time increase the user experience. User feedback helps companies navigate their product development by providing insight of their service is on the right track or if something needs to change; in order to keep the users satisfied.

This thesis will analyse the best way for companies to collect user feedback, without being intrusive and time-consuming while at the same time ensuring a high ratio of answers.

1.1.1 Telavox

Telavox AB was founded in 2003 in Lund, Sweden, as a start-up and during those 17 years has expanded to five countries with over 15000 companies as their customers. Telavox is providing business-to-business and business-to-customer unified communication as a service (UCaaS) via a cloud-based solution, combining Private Branch Exchange (PBX), telephony, chat and video solutions.

The analysis will be based on Telavox's web service product Flow and the calls made from Flow, see Figure 1.1. Flow is an adjustable PBX cloud solution, with both chat and call features that help businesses and customers alike. Both to forward calls from the company number to mobile phones and offering a system for telephone queues. To give feedback on anything regarding Flow today, a user needs to either call or e-mail Telavox. This creates a vacuum in the gathering of information where the correlation between the system values and the user experience, which gets mostly examined only when something is severely wrong. By designing a web-based rating system for Flow's user after a call has been transmitted, Telavox will have a faster and simpler way to collect their user data.

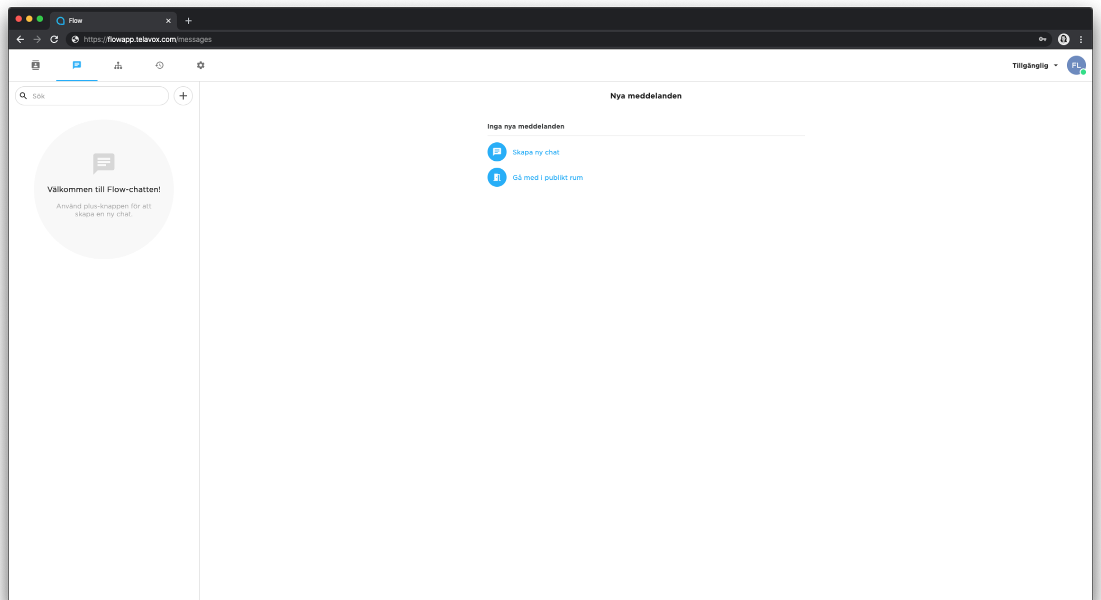

Figure 1.1: A screenshot of Flow.

1.2 Related work

Sparling and Sen (2011) investigates the four different kinds of rating scales that the base of this report focuses on; Unary, Binary, Star and Gradient. The main difference between their and our work is that theirs focuses on the same design for the rating systems but put into two different contexts, while our work focus on different design choices for one fictional scenario (Sparling and Sen 2011).

1.3 Our contribution

From our research, we have yet to find a study regarding different design choices for rating systems. We contribute with this, since we are investigating what effects animation, different colours and rating scales, among a lot of other things, have on the user.

1.4 Scope and Purpose

The main purpose of this thesis was to research how to design and implement a call quality rating system to be used for collecting data about the user experience both on incoming and outgoing Voice over IP (VoIP) calls. The data will provide Telavox with a deeper understanding of what qualifies as a "good phone call" from the user perspective, where there has only been a system perspective before. The thesis is framed by using two research questions.

- *How can different rating scales elicit different levels of user participation?*

  Different kinds of rating scales will be evaluated by a group of test people. Evaluation will be performed by rating the different rating scales in different contexts. Correlation analyses will then be used to identify which rating scales works best in which context.

- *How can different design choices elicit different levels of user satisfaction?*

  Different type of animations and design will be evaluated by the same test group as in RQ1. Users will be asked to choose the design they liked best.

1.5 Limitations

To decrease the scope of this report, some limitations were made:

- Due to the size, and the graphic charter and design directions from the company which we carry out this study for, we have decided to rule out some of the design decisions for our testing. This is because even if we would find out e.g. that a realistic design was to be more effective than a flat design, this would worsen the User Experience as a whole in the application it would be presented. Also, there are several other studies that discusses realistic versus flat design.
- For the Mid-Fi Phase, since we didn't want our user test to take too long we ruled out some systems for some combinations, e.g. a Binary rating emoji have the variation where a negative input is represented by the colour red to deduct if we could see a difference, but we did not implement a Star-scale system where we tried to extrapolate if also colour was an affecting factor.
- The evaluation of the rating system will only factor in the quality of the transmitted phone call, e.g. levels of unwanted noise and how the voice on the other end is perceived, not the connection and disconnection of the phone call.
- The thesis timeline is restricted to 20 weeks of two full-time working students. This will affect the size and complexity of the thesis.

Chapter 2

Theory

This thesis is based on several recognised processes and methods when researching and collecting the necessary data, designing and styling the rating scale prototype and dividing the workload. This chapter will describe why those processes were chosen and how they are executed efficiently.

2.1 Design and Cognitive Perception

Making use of a thorough and thought out design can be essential to stand out from competitors and to be able to grant user satisfaction. To be able to implement such a design, a lot of cognitive aspects and psychology needs to be accounted for to really accommodate what we as humans respond to biologically and to what we have experienced earlier in our lives. Disregarding the cognitive aspect of the design in an interactive system or application may severely damage the User Experience, making them seem faulty and causing stress (Norman 2002).

To create a good design, a design that both attract the user's attention and are pleasing to the eye, there are several factors to consider. Unity and contrast changes are two factors that can be considered when designing a web interface to make it as effective and pleasant as possible. Unity can be described as the harmony and balance of the design, how the different component interacts with each other. Contrast in this context is how different colours and sizes attracts the user's attention which brings life to the design (Alsudani and Casey 2009).

Studies have also shown that alongside unity and contrast, the number of compositional elements affects how aesthetically pleasing an interface appears. This means that even though an interface is designed with unity and contrast in mind, if there are too many elements to consider in the design, the user will get the impression of the interface being more unpleasant (Bauerly and Liu 2008).

Design in digital environments is something that constantly evolves. When the iPhone revolutionised the market and set a new standard for User Interfaces, skeuomorphism became the way to go when designing a new User Interface. A skeuomorphic design refers to a design that is made to mimic their real-life counterpart e.g. drop shadows under icons and shadows on buttons to mimic what we already have learned our entire lives a real button looks like. Even though a skeuomorphic design is easier for us to instinctively recognize, this is now not the current trend in an aesthetically pleasing User Interface (Duyne, Landay, and Hong 2002). Skeuomorphism has changed over the years to a more minimal flat design which now big companies such as Facebook, Google and Apple all evolved their logo into (Page 2014). An example of this can be seen in figure 2.1.

Figure 2.1: A simple logo created in a skeuomorphic and flat design.

2.2 Animation and User Experience

One important tool to aid the users in to understanding what needs to be done to fulfil a task in web design is animation. Movement in a User Interface can be crucial for a good User Experience, but to accomplish a good User Experience, the animation needs to be smooth, simple, and suitable to what the users expect to happen when performing the task (Gonzalez 1996). Studies also shows that different kinds of animation speeds elicit different kinds of physiological arousal where fast animation is more attention grabbing but slow animation speeds enhance the appeal of the website, especially when contrasted to fast animation speeds (Sundar and Kalyanaraman 2004). Animation is also proven to have a higher participation ratio in banner ads, which further strengthens the importance of animation in web design (Yoo et al. 2004).

2.3 Emojis and symbolism

With the recent growth in the usage of emojis since the 2010s, when they were added to several mobile operating systems, they have been frequently used to add visualisation to regular text messages. It was in 2010 that several Emojis were added to the Unicode Standard, thus making it easier to send emojis between different devices. A problem here arose, since it is up to each company, application or website to choose what Unicode Character represent what emoji, making it as diverse as different typefaces between devices or applications. Here confusion is not unusual since for example, sending one emoji from an Android phone will render differently on an iPhone, but also since emojis are heavily dependent on the users own interpretation (Miller et al. 2016). With the emerge of social medias such as Facebook, Instagram and Twitter, that all offer you a way of easily "praising" or reacting to an item, symbolism here has changed. An icon of a "thumb up" that was the universal sign for Okay, Affirmative, or Good, became known as a "Like" converting the thumb up meaning to something else depending on the context.

2.4 Rating Systems

2.4.1 Star- and Likert-Scale

A star scale and a Likert-scale are either generalisations or a kind of grouping of rating, where the user can choose from usually five or ten options to rate their experience or a certain product. What usually differentiates the Likert-scale from the star scale is the demand of taking a stance against a statement, even if that position is neutral. Also, a short description or text under the options is usually presented in a Likert-scale to help the user in choosing the position in comparison to the statement. This makes a star scale usually more suitable for rating the quality or experiences and the Likert-scale for questionnaires when the answer requires more complexity than good to bad. It is not unusual for a five-star scale to half-star incrementation, making it a ten-option scale.

Figure 2.2: A star scale and a Likert-scale. The star scale is showing 3 out of 5 stars and the Likert-scale is Neutral.

2.4.2 Unary

A unary rating system as seen in the left part of Figure 2.3 is a system where you only can express one kind of feeling. Facebook iconized this 'like' feature where you express your appreciation for a photo or status by clicking an icon of a hand giving a thumb up, almost like ticking a checkbox. The unary rating system does not necessarily have to describe something positive, it can for example be something negative or a flag signalling a warning of some kind (Sparling and Sen 2011).

2.4.3 Binary

A binary rating system, as seen in the right part of Figure 2.3, is what it sounds like, a system where you have two options of rating where the options usually are opposites. Usually the symbolism for this is a thumb up and thumb down (Youtube), arrows pointing up and down (StackOverflow, Reddit), or a plus or minus. Youtube switched from a star scale to a binary scale since their star scale was primarily used as a seal of approval and not as a tool of reviewing the video, since the rating options 3, 4, and 5 only made up 5% of the ratings (Sparling and Sen 2011).

Figure 2.3: Two simple illustrations depicting a unary rating system and a binary rating system, each symbolised by a thumb. The unary system is toggled, and the binary system is showing a positive input.

2.4.4 Gradient

A gradient bar, illustrated in Figure 2.4, provides the highest granularity of the scales. A gradient bar usually goes from the values zero to 100 but can vary. Studies have shown that the gradient bar is a tool which users feel provide the most accuracy during a review, but also the tool which many dislikes (Sparling and Sen 2011). The gradient bar usually has two methods of input; the gradient bar itself and a input box where it displays the bars current value. You can adjust either of these inputs and the other input corrects itself accordingly. Using the gradient bar as input, it is common to be able to both click on the desired value and to drag a handle for precise input.

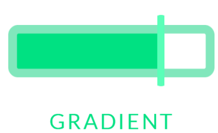

Figure 2.4: A gradient bar with a handle showing 75 out of 100.

2.4.5 Prototypes

Lo-Fi

Creating the design and interface used for software is often done before the software is fully functioning. To be able to start the design process prototypes are often used in the earlier stages of the process. Low Fidelity (Lo-Fi) means that the currently presented design will not match the final product, instead a Lo-Fi prototype is often created quickly and with easy access material like paper. The purpose of a Lo-Fi prototype is to explore the early ideas and flows of the software. An example is to simulate a computer desktop with a Lo-Fi prototype made of paper to show how the interaction between the user and the software are supposed to work (Rick et al. 2010).

Hi-Fi

A High Fidelity (Hi-Fi) prototype on the other hand is a prototype with almost identical design as the final product. A Hi-Fi prototype is often characterised by high-tech which can often be related to higher costs in both time and money. But it will give the user a more accurate view of the design and flow of the software. The interaction between the user and the software will be more realistic and because of this a Hi-Fi prototype can help the designers to catch issues with the real design before the release (Maron, Missen, and Greenberg 2014).

2.5 User-Centred Design

The purpose of a design is to make the product easy and intuitive for the user and a way to ensure this is to involve the user in the design process. Which is exactly why the term User-Centred Design (UCD) was coined by Don Norman, and was widely accepted and popular in 1986 after the publication of the book "User-Centered System Design: New Perspectives on Human-Computer Interaction", to focus on involving the user's perspective in the design. UCD includes a variety of different ways to engage the end-users, everything from the user expressing their needs to the user being a contributed participant throughout the design process (Gould and Lewis 1985) (Abras, Maloney-Krichmar, Preece, et al. 2004).

In software development UCD is recommended to be implemented with the User-Centered System Design (UCSD) process, a merged process with ideas from usability engineering process and Human-Computer Interaction design. UCSD is constructed mainly by three different phases: requirements analysis, growing software with iterative design and deployment. The focus of UCSD is the user, and therefore it is of big importance that the users are involved early on and then continuing to be consulted at every iteration made in the system process.

In the requirements analysis phase the leading issue is: understanding the main user groups and their needs. To get into the minds of the different user groups and their specific needs, user profiles and/or user cases are good ways to go. A user case builds on one test group and aligns with their needs while taking their background, skill and abilities in account. It is also important to know when and where a user is using the product. Is it for work or down time? Is it a stressful environment or relaxing? What are the user's expectations for the app? With this established it will be easier to set goals for the design and usability.

The second phase consists of three iterative loops: conceptual design, interaction design and detailed design. The user cases built in the first phase is used, along with the actual user, as the evaluation, analysis and redesign iterations go on. The iterations will help the design process to go from overview to detailed.

The third and last phase of the UCSD is the deployment phase. In this phase the design and software are fully developed, and the only thing left is to make sure the user knows how to work the product by providing proper instructions, manuals and training. This phase differs depending on the type and size of the organisation, but the important thing is to include this phase in the timeline and scheduling from the start. Another tip is including the deployment phase in some of the iterations to make it easier to present the software in the earlier stages because of the prototype having less features than the end product (Göransson, Gulliksen, and Boivie 2003).

2.6 Feedback

When interacting with other people, humans get constant feedback in the form of facial expressions, words and body languages. In a research made about gaming the result showed immediate feedback, among other things, were irreplaceable in terms of user engagement. By using feedback as a way to communicate about actions and results, software simulates human-to-human contact and makes the user feel seen. The feedback given from a system can take many different forms some of the most common being visual or auditory, giving the user a sense of a common environment (O'Brien and Toms 2008).

The same research also shows the importance of feedback when trying to keep the users engaged throughout tasks. When a task has begun, feedback plays an essential role in keeping that focus. As a tactic to make the user feel like a part of the interaction, and connected to the system, the system updates the user on processes, actions and results. Together with a pleasing interaction, design feedback helps with keeping the users invested (O'Brien and Toms 2008).

2.7 Evaluation methods and processes

2.7.1 Brainstorming

The main idea of brainstorming is a creative environment without judgement or criticism for people to express their ideas, big as well as small, about a specific question. One of the two principles of brainstorming is "Reach for quantity", meaning to enhance people's freethinking and spontaneous ideas and by doing so increase the chance of good ideas. The other principle is "Deferring judgement". Not only will it help people to think more freely but it can also help breeding more outside-the-box ideas, expanding the overview and hopefully resulting in better solutions. All ideas discussed will be written down for an evaluation later on.

Brainstorming is a group activity that Alex F. Osborn, the man that first developed brainstorming as a creative method, envisioned this method for groups as large as 12 participants. The brainstorming session would be beneficial by mixing the participants expertise, age and ethnicity for a better diversity in the discussions. After 1939 when Osborn first came up with brainstorming the method has evolved and different versions of brainstorming exists and is being used today (Wikipedia 2019).

2.7.2 Data Gathering

There are two different data gathering processes, quantitative and qualitative data. The two differs with how they measure data, with quantitative data gathering how well test users perform and qualitative data gathering why a test user performs as they do. Two other categories for dividing data are subjective and objective data. These categories help the developer with how to interpret the data gathered, where subjective data is based on the users opinion on the performance and objective data is based on the result after the performance (Rubin and Chisnell 2008). Combined they form four important categories of data gathering:

- Quantitative objective - For example measuring time or clicks.
- Quantitative subjective - For example opinions on the performance.
- Qualitative objective - For example observations.
- Qualitative subjective - For example open questions.

2.7.3 Cognitive Walk-through

Cognitive Walk-through (CW) is a theory created by Norman based on task execution in seven steps: goal establishing, intention formulating, actions specifying, executing, perceiving, interpreting and evaluating (Mahatody, Sagar, and Kolski 2010). This process has been evolving with the digitalising age and from that a new theory has been formed; Cognitive Walk-through for the Web (CWW).

In similar steps as CW uses the CWW simulates the users performing tasks by assuming the user to be goal driven. In addition, the CWW focuses only on websites and web-based solutions' design and navigation. In order for the CWW to be better adapted for web solutions three main features are considered. A contextually description of the user goals is the first step, and this is important for the users' understanding of the goals and their motivation. The second step is regarding the interaction with actions, for example clicking on a button or link. Parsing the new page into components and focusing on the correct component and then choosing the right part of the component to interact with. The third and last feature that differs from the original CW is that in CWW the performer works on one web page at time, making the organisation more adapted to web-based solutions. This process is then used in every iteration of the design development (Blackmon et al. 2002).

2.8 Psychological influence

Context effects regard the effect on how the surrounding elements can affect the user. Depending on what context something is placed in, the opinion changes and or is weighted. As Tourangeau and Rasinski wrote regarding Interpretation of Unfamiliar Issues, they saw a strong variation in the results depending on how the items in the questionnaire were grouped or if they were scattered. Context can also suggest a standard to which subsequent items are judged when being compared to each other. This means that not only is an item affected by what other items it is grouped with, but also since the prior items may provide a basis from where the next item is evaluated against (Tourangeau and Rasinski 1988).

Anchoring is a psychological effect that describes the influence of an initial value on answers. The problem has its roots in insufficient adjustment to the initial value. An example of this is a study where the subjects were asked the percentage of African countries in the U.N. and was given an initial percentage, then asked if the initial value was too high or too low, and then to estimate the percentage. This study showed for estimation of quantities in percentage, the median between groups increase drastically. The first group were given the initial value 65% and the second was given 10%, which resulted in a median of 45 and 25 respectively (Tversky and Kahneman 1974).

Chapter 3

Method

This work was divided into four main phases.

- In the first phase (Investigation Phase) we gathered our theory and background knowledge in a literature study regarding basic aspects of rating systems, interactivity and cognition, so that we had enough knowledge to plan and design suitable rating system that we could extract sensible conclusions from. During the literature study the scope was narrowed down with brainstorming sessions, a cognitive walk-through for our web solution, determination of requirements and the results found in the literature study, with both the university and the company weighing in.
- In the second phase (Mid-Fi Phase) the development of the testing application and the user tests were the main focus. A quick Lo-Fi prototype of the different rating scales was created as sketches after some discussions, CWW and brainstorming sessions before we started the developing process. The digital testing application was developed as a web based React application using Redux as a state container, with mongoDB as our database. The digital testing application gathered not only the different assessment values, but also measured the time spent on each of the rating systems, to include all the four main categories of data gathering. This kind of prototyping was chosen because of the realism and likeness between the digital testing application's design and the design of the end product. We felt a paper prototype would not give the testers a realistic overview of either the design of the rating scales nor the digital application the rating scale was to be implemented in. For the Mid-Fi

Phase, we did two iterations of our user tests. The first iteration was a trial run on a small test group with seven people. The second iteration started with revising the design of the digital testing application and rating scales, after considering the feedback from the small first iteration, the test was carried out on 26 people. The rating scales were divided into two main categories, one with our baseline rating scales and the second with our more experimental rating scales.

- In the third phase (Hi-Fi Phase) we evaluated our test results based on the ratings given in the testing application's three assessments as well as our observations and notes taken during the tests. Both the assessment results, time results and the observations were quantified with the help of Excel for an easier overview of the most liked and disliked scale and why. Mapping was discussed and a prototype in the application was presented with the main purpose of discussing how the placement of the scale affects the quantity and the quality of the ratings. A new design iteration was made where all the highest scored rating scales where modified based on the result. These final design choices were presented for a small test group consisting of five Telavox employees, based on their requirements for the usage the rating system, in Telavox's product Flow. The test persons were asked to rank the three rating scales and also to give their opinion on what they thought were the main pros and cons of the rating scales.
- The fourth phase (Implementation Phase) consisted of the final design iteration of the single rating system left, where the result from the Hi-Fi Phase was taken in to consideration and applied to the winning rating scale. After the creation of the final rating system implementation into Flow followed. Besides the UX part of the implementation the back-end part of the implementation contained connecting the given rating to an API point making Telavox able to store and use the data.

Chapter 4

Investigation Phase

Concurrently as we did our literature search, we also brainstormed a lot regarding all of our design choices as we dove deeper into the theory of rating systems, cognition and symbolism. We discussed what we had learned after each article and how we could design a rating system to test the specific theory, e.g. intuitive animations.

4.1 Design Choices

To confirm our theories about all the four main rating scale systems we decided to combine all the systems with each other, with a few exceptions.

Regarding the Likert implementation, we decided to exclude it from our testing since Likert-scales usually are used to describe statements rather than good/bad quality, which is the only kind of feedback we are aiming to be collecting from the call. However, we will be testing the psychological effects of being presented different kinds of text choices, but those are not applicable to a direct scale but instead problematize how the user reacts when presented different options for possible outcome qualities of a call.

To guide the user into understanding how to use rating system more easily, we decided to use only a thumb up for the unary base, Figure 4.1, a thumb up and down for the binary base, Figure 4.2.

Figure 4.1: Unary neutral.

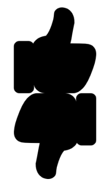

Figure 4.2: Binary neutral.

The star scale base as five stars in a row, Figure 4.3, and a rounded off bar for the gradient bar base, Figure 4.4. Leaving us with four different scale bases to combined, making it a total of 16 rating scale prototypes to test in our baseline testing.

Figure 4.3: Star neutral.

Figure 4.4: Gradient neutral.

Our 16 different baseline versions can be seen in table 4.1. To help illustrate the combinations, four of them can be seen in the following four figures, US in Figure 4.5, BN in Figure 4.6, SG in Figure 4.7 and GU in 4.8.

Table 4.1: Our combined baseline Mid-Fi rating scale prototypes.

No.	Base	Alteration	Code
1	Unary	None	UN
2	Unary	Binary	UB
3	Unary	Star	US
4	Unary	Gradient	UG
5	Binary	None	BN
6	Binary	Unary	BU
7	Binary	Star	BS
8	Binary	Gradient	BG
9	Star	None	SN
10	Star	Unary	SU
11	Star	Binary	SB
12	Star	Gradient	SG
13	Gradient	None	GN
14	Gradient	Unary	GU
15	Gradient	Binary	GB
16	Gradient	Star	GS

Figure 4.5: The Unary thumb as base and the five-star steps as alteration.

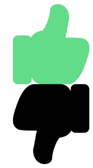

Figure 4.6: Binary thumbs as base and Binary 50 or 100 as alteration.

Figure 4.7: Star as the base and the 0-100 gradient steps as alteration.

Figure 4.8: Gradient as base and the one unary step as alteration.

From our theories we also wanted to investigate more experimental rating scale prototypes and how different design choices for example a basic Star scale versus five different kinds of Emojis influenced the test participant. We also wanted to test how different kinds of animation affected the test participant for example if the star that was clicked enlarges for a quick moment or if it 'filled up' gradually. Added together we had 25 different rating scale prototypes we wanted to evaluate.

Our 9 different experimental versions can be seen in table 4.2.

Table 4.2: Our experimental Mid-Fi rating scale prototypes.

No.	Base	Alteration	Code
1	Binary	Arrows	BA
2	Binary	Emoji	BE
3	Star	Choices	SCh
4	Star	Confirmation	SCo
5	Star	Emoji	SE
6	Star	Half stars	SH
7	Star	Stroke	SSt
8	Star	Suggest	SSu
9	Gradient	Slider	GSl

4.2 Digital Testing Application

As described in the chapter Method, we wanted to develop a digital testing application, see Figure 4.9, for our rating scale prototypes to simulate the realness of the digital environment the final rating scale prototype would be implemented in. Based on studies on aesthetic appeal on the numbers of compositional elements and studies regarding the fact that aesthetically pleasing design works better it was also in our interest to design our testing application with as few components as possible without compromising the test itself.

To gather quantitative subjective data, we choose to include three assessments in our digital testing application. For every rating scale presented in the testing application the test participant were to appraise three assessment regarding expression, design and simplicity. These three assessments were:

- Assessment 1: Grading quality. Did you feel satisfied with the way your rating expressed your feelings for the experience?
- Assessment 2: User Satisfaction. How aesthetically pleasing was the user experience? How much did you enjoy using the rating scale?
- Assessment 3: Ease of use. How easy is the rating scale to interact with?

The assessments were designed to take in a grade between 0-100 where 0 was the lowest and 100 was the highest. This because the test participant should be able to express exactly how they felt about all the questions, while not bearing any resemblance to any other of the rating scale prototypes.

To gather quantitative objective data, as well as help the test participants to decide of the simplicity of the rating system prototype, we choose to include two assignments.

- Assignment 1: Give the highest possible score of the rating scale prototype.
- Assignment 2: Give the lowest possible score of the rating scale prototype.

When an assignment had been fulfilled a checkbox was ticked, to indicate the completion of the assignment. A timestamp with the time of the completed assignment was then also saved for us to later evaluate. We also took a timestamp when a test participant was done with the evaluation of the rating scale prototype to measure the total time spent on the prototype.

To be as thorough as possible when testing our rating systems we wanted to use all of the four main data gathering choices. To be qualitative objective and qualitative subjective we decided writing down our observations during the test and asked a few open questions to the test participants after they had tested all of our rating scales.

When all the assignments were completed and assessments were answered, the system would allow the tester to continue which would present a new rating scale to be evaluated. A sketch of how the testing application was envisioned can be seen in Figure 4.9.

Figure 4.9: A first mock-up of The Digital Test Application.

Chapter 5

Mid-Fi Phase

The Mid-Fi Phase started with the development of our digital testing application and the production of our first 25 different rating scale prototypes. After a functional test application was up and running, we started with our first test iteration.

5.1 Test Iteration 1

5.1.1 Test Participants

The first iteration was conducted on a small test group of seven people that all have been involved in the development of the testing application, rating scales and were habitual users of Flow. These people were ranging from other co-workers that have been interested and discussed the system with us, to our supervisors and the company's Head of UX. The test group was put together to act as an trial run to properly test both the digital test application's features and give a quick evaluation of the 25 rating scales in Table 4.1 and Table 4.2.

5.1.2 Execution

We introduced the digital testing application and asked the testers to think out loud and give us feedback not only regarding the rating scales, but also if they had any comments regarding the digital testing application afterwards. We observed both the application closely when the participants were taking the test, to discover any possible bugs that they themselves might not notice or comment on, and the participants themselves, to gather their facial expressions and clicking patterns. We, as test conductors, didn't engage with the test participant during the test, even when questions were asked, since we wanted to get genuine data of how easy the test participant found the different rating scales. After the test was done, we asked the test participant open questions.

5.1.3 Result of Iteration 1

Result of the Assessments

The calculation (in this case for the Grading Quality) is described by the following equation:

$$X_{Rel.GQscore} = X_{GQscore} - \sum_{k}^{N} \frac{k_{GQscore}}{N} \tag{5.1}$$

$$GradingQualityScore = \sum^{M} \frac{X_{Rel.GQscore}}{M} \tag{5.2}$$

where N are the number of rating systems, $k_{GQscore}$ the Grading Quality score for rating system k, $X_{GQscore}$ the Grading Quality score for rating system X, which then gives us the relative positive or negative score $X_{Rel.GQscore}$. All of the relative Grading Quality scores are then summarised from all test persons M. This formula was then repeated for the other two assessments.

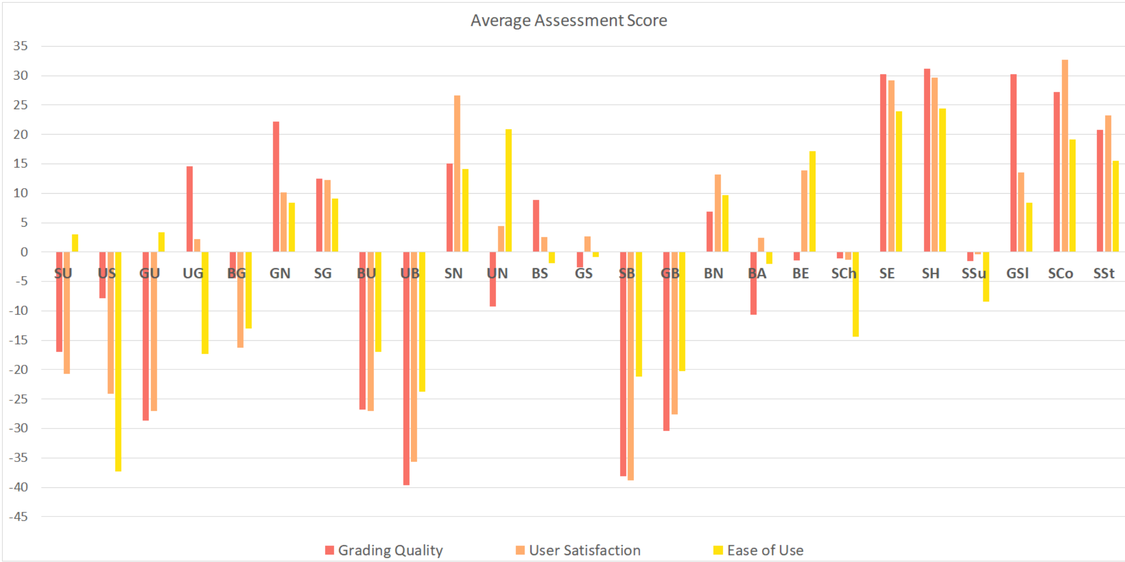

Figure 5.1: The score from assessments Grading Quality, User Satisfaction and Ease of Use depicted on the Y-axis for each rating prototype.

From each of the three different assessment an average score was first calculated based on all the assessment scores gained during the first test iteration, as described in equation 5.1 and 5.2.

The assessment scores in Figure 5.1 are showed in three different stacks, red for Grading Quality, orange for User Satisfaction and yellow for Ease of Use. Showing us that the SH prototype gained the highest Grading Quality score, SCo prototype the highest User Satisfaction score and the SH prototype the highest Ease of Use score.

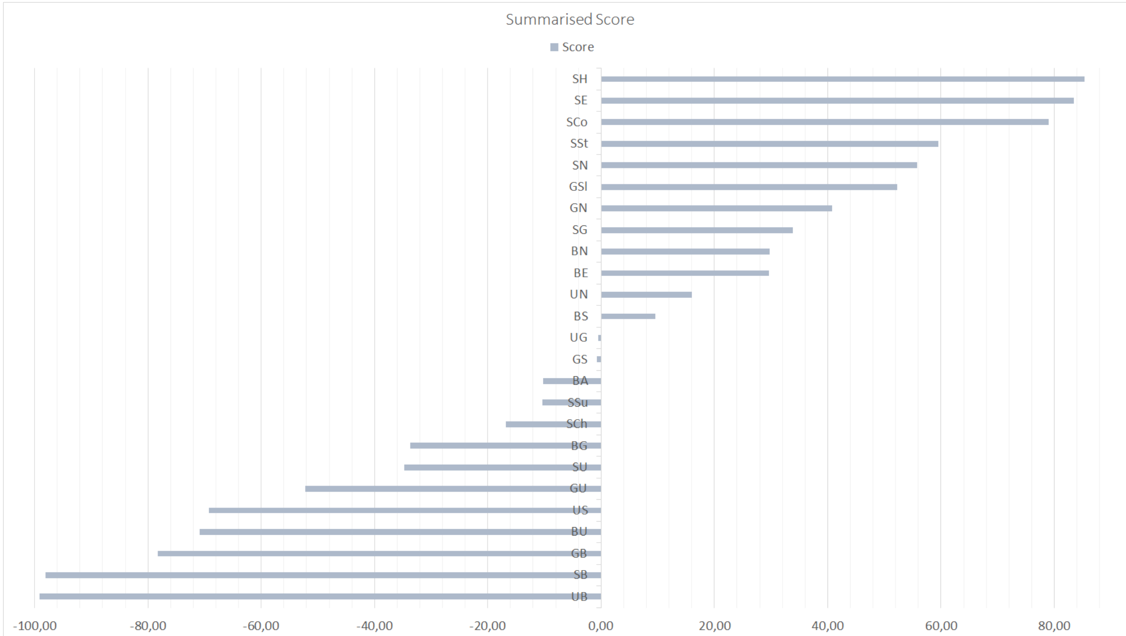

Figure 5.2: The summarised scores from assessments Grading Quality, User Satisfaction and Ease of Use depicted on the X-axis for each rating prototype.

The average score was calculated with the same equation 5.1, only this time the different scores from the different assessments were then summarised together, giving the Figure 5.2 a range between -100 to 80. Theoretically, if all testers were to rate everything zero except one rating system, the total score for a system could then approach 300, respectively -300 if it were the other way around. This would of course mean that all of the other systems would approach zero since that is their average given rating. For each point over the average score the rating scale prototype increases with one point in the graph, and similarly for each point below the average score the point in the graph decreases. These scores were then summarised to be read in Figure 5.2. Topping the graph is the SH prototype with 87 point.

Both graphs can be found in Appendix A, Figures A.1 and A.2, as larger figures.

Result of Timestamps

Unfortunately the timestamps were not saved during this iteration due to problems with the database.

Result of notes and observations

The notes were quantified by analysing the comments the seven test participants said and expressed during the test.

Almost all of the test participants commented that it was hard to keep track of their previous inserted scores, making it hard to be impartial and concrete with the assessment scores. We also noticed that even though different rating scale bases were used some of the prototypes gave the same reaction and comments. A few comments were made about the speed of the different animations.

5.1.4 Conclusions from Iteration 1

We learned a lot from our first test iteration and trail run. We decided to streamline our rating systems by removing those that were too non-intuitive and were deemed unnecessary to find out any meaningful conclusions compared to what the other tests contributed with. We were not satisfied with the animation in SSt and therefore we choose to delete it.

We also felt 25 rating scale prototypes were too many. The rating systems deleted can be seen in Table 5.1.

Table 5.1: The deleted rating scale combinations.

No.	Base	Alteration	Code
1	Star	Binary	SB
2	Binary	Unary	BU
3	Star	Unary	SU
4	Binary	Star	BS
5	Star	Stroke	SSt

We made some small tweaks to the more experimental scales, where a more rating system unique behaviour heightened the experience for the rating systems, as well as helped us to extrapolate some design aspects more distinctively. The order of the systems was changed to be more varied and improved some of the describing texts for a easier understanding of the systems.

We think that the general high scores, Figures 5.2 and 5.1, in this test iteration are a product of the test group being small and somewhat involved in our earlier design process. This made it easier for them to decide which prototype they liked and which they didn't like and therefore giving more extreme assessment scores.

5.2 Test Iteration 2

5.2.1 Test Participants

Since our second test iteration was the foundation of our study the test group needed to be bigger to get a than the first test iteration. The goal was to get our rating scale prototypes tested and evaluated by 30 people, but in the end only 26 of these testers were able to participate. The test participants mostly consisted of students and employees at Telavox. The students that we chose for the tests were chosen to have as diverse line of study as possible to get as much of a mix of opinions as possible. Here the ages ranged from 21 up to 27. For the employees at Telavox, we tried to gather as many "Advisors" as possible, since the Advisor position at Telavox is the one that is most in touch with the everyday user through phone contact. For the employees of Telavox the ages ranged from 20 up to 39.

5.2.2 Improvements

These are the improvements that were made from the conclusions made during the first test iteration.

Design Improvements

The decision to add Star Four for this iteration was made since some test participants commented on the star systems that the middle star was a neutral choice. Because of this we designed a star system without the middle star to force the user to take a stance, much like in a Likert-scale. Another implementation worth discussing would be how the result would have varied if we had implemented six stars instead of four. We changed the colour of the negative emoji for BE to be red to enhance the visualisation of the emoji giving a negative review when toggled, and to examine how colour changes appeared after having all the other tests consisting of the same green colour when toggled. The animation's speed for the colour fill when clicking an element was also increased due to the feedback we received, giving all actions a more natural response. All the 21 rating scale prototypes tested in the second test iteration can be found in Table 5.2.

Table 5.2: Our Mid-Fi rating scale prototypes for the second test iteration.

No.	Base	Alteration	Code
1	Unary	None	UN
2	Unary	Binary	UB
3	Unary	Star	US
4	Unary	Gradient	UG
5	Binary	None	BN
6	Binary	Gradient	BG
7	Star	None	SN
8	Star	Gradient	SG
9	Gradient	None	GN
10	Gradient	Unary	GU
11	Gradient	Binary	GB
12	Gradient	Star	GS
13	Binary	Arrows	BA
14	Binary	Emoji	BE
15	Star	Choices	SCh
16	Star	Confirmation	SCo
17	Star	Emoji	SE
18	Star	Four	SF
19	Star	Half stars	SH
20	Star	Suggest	SSu
21	Gradient	Slider	GSl

Performance Improvements

During the first test iteration we noticed that most of the nine more experimental rating scale prototypes were not compatible with our desire to measure the time of the highest and lowest possible scores. Mostly because some of the experimental prototypes were designed in a way making it hard to decide what was the highest score and what was the lowest score. Therefore in the last part of the test we disabled the maximum and minimum goals in the digital test application. After improving our digital test application to get rid of some bugs we decided on two of the open questions we had tried out during the first iteration. The following questions were chosen as we felt they were open and would still give us valuable data.

1. Which one of the rating scales are you most likely to use again?
2. If you could combine features from the rating scales you have tested today, or at any other time, how would you design your dream rating system?

The answers to the questions, together with our observations and notes, were written down in an Excel sheet for later evaluation. We also thought out a user case for the test participants to imagine before they started evaluating the last nine prototypes, the more experimental ones.

- User case: You just got a phone call from a friend saying: 'I will meet you in five minutes', and then hung up.

This was done because of our goal to implement one of the prototypes at Telavox and therefore the user case was telephone based to provide a user context for the rating prototypes. We also wanted a neutral user case to prevent emotional response while rating the call quality.

5.2.3 Execution

We started out with introducing ourselves and then our digital test application, see Figure 5.3.

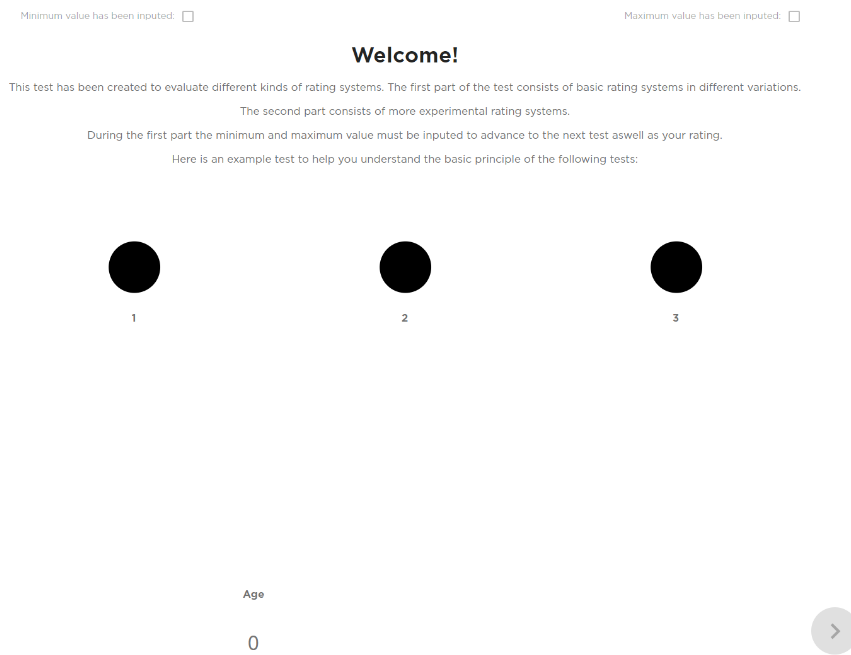

Figure 5.3: The Digital Test Application welcoming screen that greets the user when they first start the test.

First introducing the purpose of the test, to rate different kinds of rating systems, followed by explaining that this is to be done using the three assessments as well as to think out loud for us to take notes. We explained how each of the rating systems are able to give a maximum and minimum score, see Figure 5.4, which was something demanded of the test participant to utilise before being able to move on to the next rating scale.

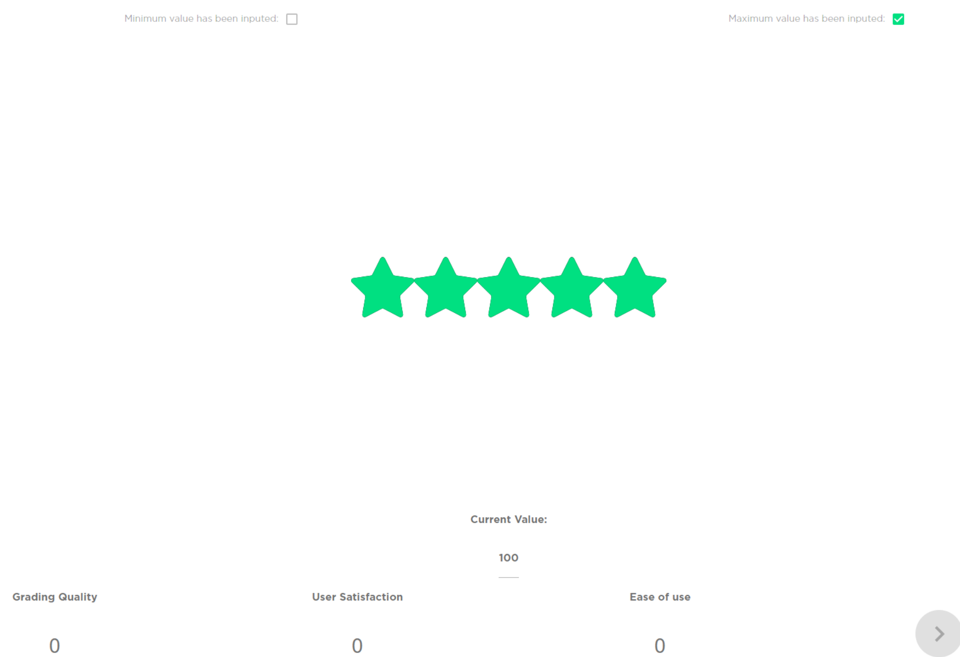

Figure 5.4: The Digital Test Application with the Maximum Value input for a star based system. No assessments has been filled out yet (Symbolised by a dark grey zero as a placeholder.)

We asked each test participant to first try to complete the assignments then fill out the assessments to make sure that the tester had understood the rating system fully before judging it. We also explained that a user case would be given, for a more accurate scenario for the test participants to insert themselves in. The test ended with us asking our two open questions described earlier.

5.2.4 Result from Iteration 2

Result of the Assessments

Figure 5.5: The score from assessments Grading Quality, User Satisfaction and Ease of Use depicted on the Y-axis for each rating prototype.

In Figure 5.5, the average score was calculated in the same way as before with equation 5.3 and 5.4:

$$X_{Rel.GQscore} = X_{GQscore} - \sum_{k}^{N} \frac{k_{GQscore}}{N} \tag{5.3}$$

$$GradingQualityScore = \sum^{M} \frac{X_{Rel.GQscore}}{M} \tag{5.4}$$

The assessments are shown with the three different bars; Orange for Grading Quality, green for User Satisfaction and blue for Ease of Use. Showing us that the GN prototype gained the highest Grading Quality score, SCo prototype the highest User Satisfaction score and the SN prototype the highest Ease of Use score.

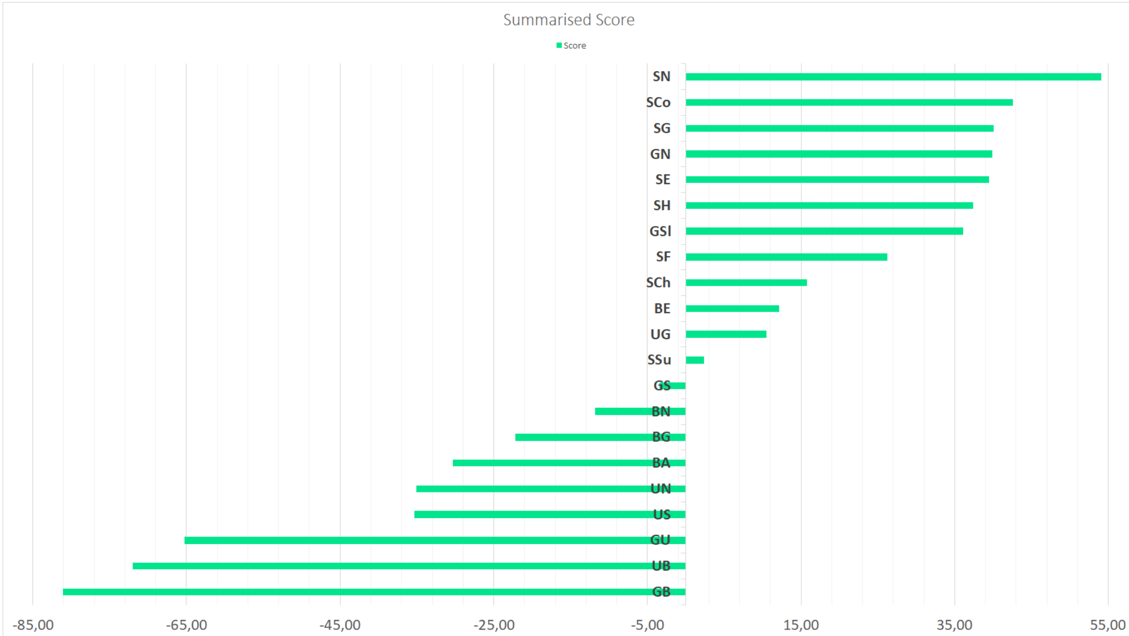

Figure 5.6: The summarised scores from assessments Grading Quality, User Satisfaction and Ease of Use depicted on the X-axis for each rating prototype.

Equation 5.1 was used for the calculation of the average. For each point over the average score the rating scale prototype increases with one point in the graph, and similarly for each point below the average score the point in the graph decreases. These scores were then summarised to be read in Figure 5.6. Topping the graph is the SN prototype with 54 point.

Both graphs can be found in Appendix B, Figures B.2 and B.1, as larger figures.

Result of Timestamps

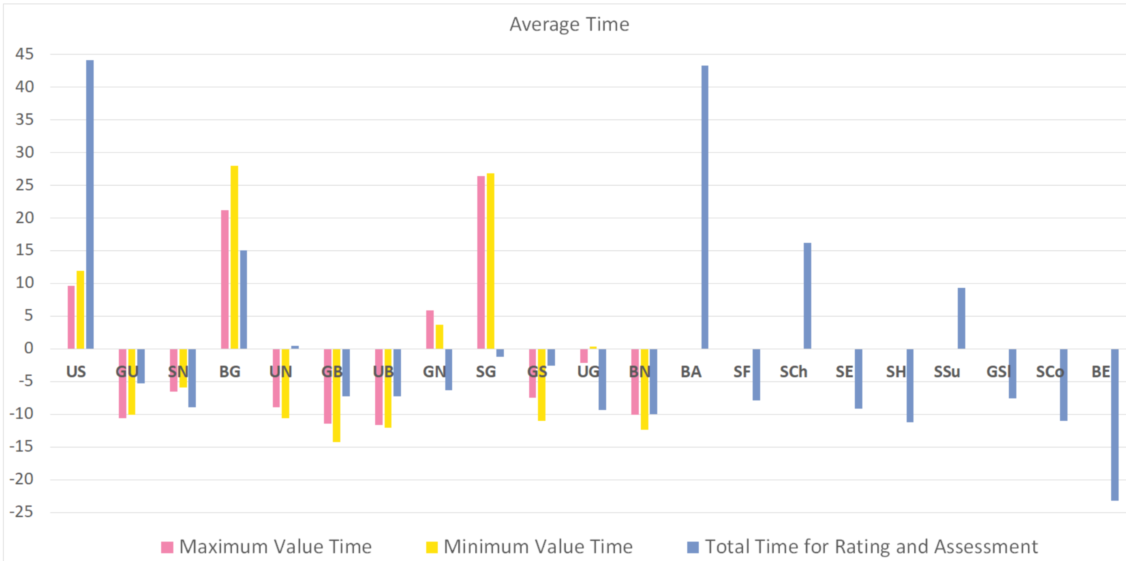

Figure 5.7: The Y-axis shows the time average score for test participants to achieve the maximum and minimum value, as well as the total time spent on the system. The X-axis shows for which rating prototype.

The time graph in Figure 5.7 was calculated the same way as the two graphs showing the assessment result; an average time for the test participant to score the highest possible score, one for the lowest possible score and then the average total time a test participant would spend on the rating scale prototype. For every second longer than the average time it took the test participant to finish the two assignments one second was added in the graph. As can be seen, the US prototype together with the BA prototype had the highest total time, which in the BA case can be excused as the user case being explained. The SG had the highest time for the maximum value, and a close second on the minimum value, which the BG prototype had highest.

Since the nine more experimental Mid-Fi prototypes (the nine last ones in the graph) didn't have a maximum or minimum value, the only time showed in the graph is the total time the test participant spent testing the prototype.

Result of notes and observations

The notes were quantified by analysing the commentary run by the 26 test participants during the test and then clustered into single words. The clustered words were evaluated on a scale -1, 0 or 1 based on the nature of the response, negative, natural/no comment or positive.

Table 5.3: The result of the observations of our prototypes.

No.	Code	Score
01	US	-14
02	GU	-20
03	SN	14
04	BG	-20
05	UN	-14
06	GB	-21
07	UB	-17
08	GN	12
09	SG	3
10	GS	-12
11	UG	-7
12	BN	-5

Table 5.4: Continuation of the result of the observations of our prototypes.

No.	Code	Score
13	BA	-10
14	SF	17
15	SCh	8
16	SE	18
17	SH	1
18	SSu	-15
19	GSl	13
20	SCo	23
21	BE	9

The three systems with the highest scores are:

1. SCo with 23 points
2. SE with 18 points
3. SF with 17 points

The most commonly used clustered words for the top three liked system were:

1. SCo - Encouraging, Nice Design and Great
2. SE - Intuitive, Easy to Understand and Fun
3. SF - Intuitive, Simple and Misses the fifth star

The least liked system was:

- GB with -21 points.

The most commonly used clustered words for the least liked system were:

- GB - Weird, Unspecific and Non-Intuitive

Result of the first question

Which one of the systems would you like to use again?

Table 5.5: The result of the first question.

Code	Number of Votes
SN	6
SCo	5
SG	4
SE	4
SH	3
GN	2
SCh	2
GSl	2
BE	1

Out of the 21 presented system nine of them were mentioned when the test participants were asked the first follow-up question. Six out of the 26 test participants said they liked the stars with no alteration the best, five said they liked the system with the stars and the feedback and on third place with four votes each are the five emojis and the stars with the gradient feature. Some people mentioned two systems that they liked equally as much. Only two of the nine systems were from the Mid-Fi baseline prototypes resulting in seven systems from the more experimental prototypes.

Result of the second question

If you could combine features from the systems you have tested today, or other times, how would your dream rating system look like?

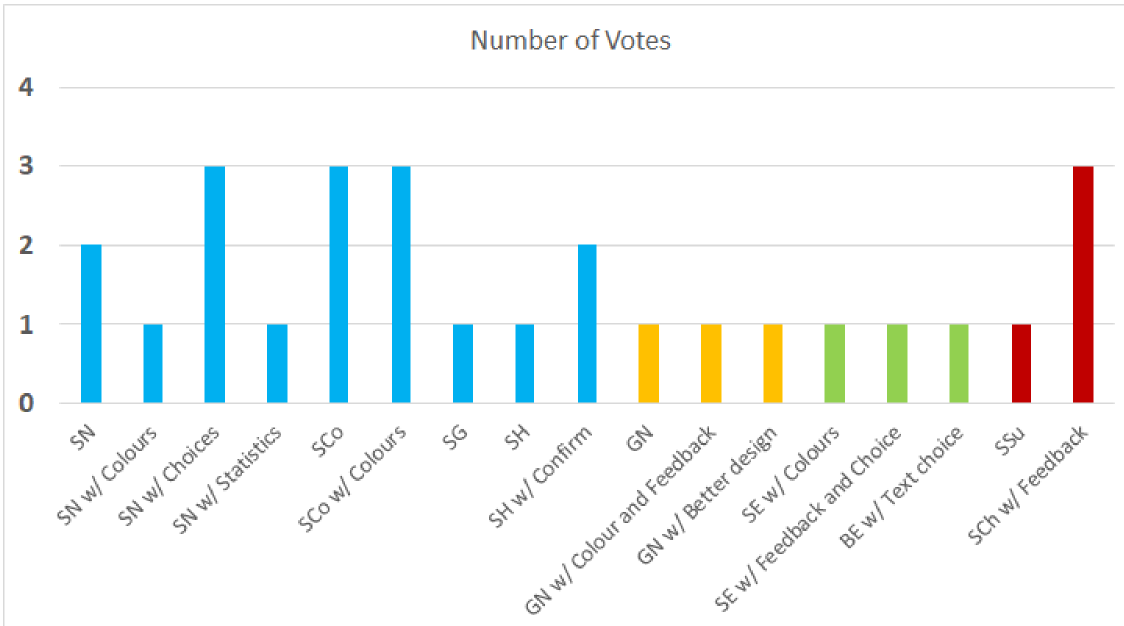

Figure 5.8: Graph showing the number of votes from the second question. The blue colour represents Star system combinations, yellow the Gradient system combinations, green the Emoji system combinations and red the other combination of systems

As we can see from Figure 5.8, 17 of the test participants answered with a combination of the Star based system (Blue). The Gradient (Yellow) and Emoji (Green) based systems each had three test participants answering they would like some kind of combination of those, and four answered other system combinations all together (Red).

5.2.5 Conclusions from Iteration 2

We saw a strong trend in positivity towards the Star-based systems. A system that stands out in our result is the standard star scale (SN), which we concluded was partly due to an anchoring effect for the system, since it was the first real intuitive system that the tester encountered.

On the opposite side, we discussed that the second highest rated system, SCo, suffered from diminishing returns, since most of the rating systems previous SCo all received positive responses. From our interview notes, we confirmed this theory, since most of the comments regarding SN were purely positive with comments like "Easy to use" or "Classic", which was also strengthened by the result of the time graph. While SCo received appraisal for making it feel like the user had been heard and that the rating mattered to the client and that the test participant had contributed by rating, but also comments about the grading quality or the colour choices, which SN didn't receive. But with SCo being one of the most popular one when the first question was asked our theory, where system feedback increases user engagement and by so increases the chance of people using the rating system again, is strengthened.

This result was close to our hypothesis that in general that five stars are quickly associated with rating, making it easier for the test participants to understand and use the star prototypes.

The smileys were popular; the SE scored the second place based on our observations in Tables 5.3 and 5.4 and BE the fifth place. They were quickly recognised in a rating environment as well, with the different being more associated with feelings and atmosphere. Many of the first reactions on the SE system were about airports, pharmacies and stores, rating the entire experience or feeling rather than more objective items. Some participants even commented that the smileys were too childish and didn't fit for the purpose of expressing quality or functionality. Based on both comments like this, and at the request of Telavox, systems with smileys will not be researched further.

Besides the Star-based systems the Gradient-based systems were popular as can be seen both in Figures 5.6 and 5.5, and Table 5.3. The two gradient-based prototypes GN and GS were the only ones besides SN to achieve positive ratings from the baseline prototypes. And the only slider of our more experimental prototypes, GSl, scored forth most popular from our observations in Table 5.4. The participants liked being able to express themselves with specific numbers and recognised the sliders from a rating point of view. However, the result can be misguiding since a lot of test participants gave the gradient based prototype a high score when asked to evaluate the grading quality, associating many options as good grading quality. There were also a lot of negative comments about the gradient scales being too specific and too detailed. We are also aware of time pressure being a parameter not included in the digital test application. So even if using a gradient rating scale prototype once or twice can be encourage from our digital test application, the 10th time being asked to rate with a slider might not be as appreciated. Worsening the desired result of a quick and frequently used rating scale that we and Telavox would like it to be.

Neither the unary-based nor the binary-based systems were well received by the 26 test participants. Some recognised and liked the thumb symbol, and the fact that a choice was forced to be taken. But the majority of both the criteria results were negative, and only one person wanted to use any of the unary or binary-based systems again, as seen in table 5.5. This was also strengthened by the theory and our hypothesis that people wants to express themselves more accurately when asked about quality than neither the unary nor the binary systems could provide.

For the SCh system to be presented to Telavox as a candidate the text alternatives needs to be extended with more alternatives and more objective words. For example, the alternative 'Uneven' was not a word associated with phone calls and the term 'Unclear' was commented for that particular alternative. The mapping of the different suggestions was also commented on, saying it affects the way the user will think back on the call. However, the suggestions were well received and a lot of people liked the association with the phone call, forcing the users to engage and comment on the quality.

One of the most disliked systems was SSu, which was a system taking form with input from Telavox and their requests. The intention of the system being based on the combined values of Telavox's server status, the connection between devices and packages lost between devices. This would in turn lead to Telavox offering the user a rating on how they thought the call's quality had been, based on these three parameters, and the user having the opinion to agree or disagree. The hypothesis about this system being about increasing the user satisfaction after a phone call with bad quality, making the user realise Telavox already being aware of the problems and by extension working on improving them. Making the user feel seen and heard even before sending their rating. Even so, many of the participants misinterpreted the complimenting text 'Calculated rating: 4.2' as a mean value of other participants earlier ratings, feeling it was influencing their own experience. There was also confusion about the score not matching the rating inputted from the participant, and frustration that '4.2' wasn't a score the participants were able to input with the star scale that was presented. Our conclusion is that this implementation was a design mistake on our part, which is why we took the decision to go ahead with this system for our next stage, just with an improved design to see if the conclusions were correct.

Other comments about the systems was the lack of colour, which was showed in the result table of the second question, Figure 5.8, where the colour was a popular feature to combined the different systems with. Especially the colours red, for negative ratings, yellow, for mediocre ratings, and green for the positive ratings.

Chapter 6

Hi-Fi Phase

6.1 Purpose

After the results and conclusions from the Mid-Fi Phase the rating scale prototypes were narrowed down to three top prototypes, which were redesigned to contain the best parts of each similar tests but with some variations.

The main goal of the new prototypes was that they should contain as few clicks as possible, and be able to provide Telavox with a lot of information while having a high answer ratio.

6.2 Testing Platform

Being the last stage of the test iterations, the testing platform for the Hi-Fi phase was performed within Telavox's web application Flow. During this stage we also experimented with different placements and access points of the rating systems within the test environment to find the easiest point of access to maximise user participation.

6.3 Test participants

For this phase we did a smaller test round with five people. We handpicked one test person, Head of the Advisor department at Telavox, who had not tested any of our prototypes before. This because we then got a new set of eyes on the design and also because he as an Advisor would use this feature the most. We also tested our last three prototypes on The Head of UX at Telavox, one UX Team Lead, one web application front-end developer and one back-end developer. All who had already tested previous prototypes. The age range was from 27 up to 39.

6.4 Design Choices

To create a more Telavox streamlined rating system we choose not to continue experimenting with colours. Even if this was one of the more popular combinations, both we and Telavox wanted a non-intrusive rating system which we felt we would achieve with a more minimal colour scheme.

Therefore different placements and access points were examined during this phase. The access points were on three main places: under the contact information when a specific call has been clicked, a small button next to the call history-item and as a button on the contact information. The two latter access points then showed only the rating systems. The two latter were also discarded before the testing phase after consulting with Telavox's Head of UX, since the button next to the call history-item was considered to be of too valuable space and hurtful to the minimal design that the call history-item has, and since the button on the contact information was harder to access than under the contact information.

We did choose to further test two different kinds of star based systems, because of the popularity, and all three of the systems we implemented had a feedback confirmation after the rating was done.

The choice and statistic parts of the systems was implemented and further tested because we felt like our previous prototypes did not do these options justice. And those are two combinations that would give Telavox additional information about the quality of the call. Our three prototypes for the Hi-Fi Phase was:

- Choices with foldout
- Star with statistics
- Star with statistics and foldout

6.4.1 Choices with foldout

A prototype quite similar to the Mid-Fi Phase's Star with Choices, but with simpler layout, having all the choices on the same row, see Figure 6.1.

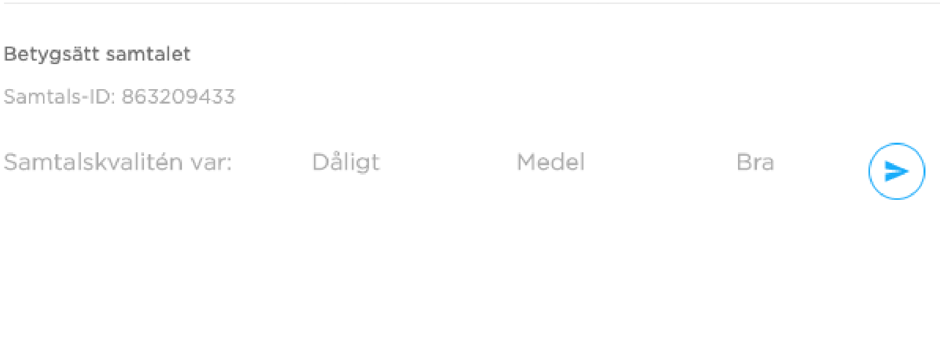

Figure 6.1: Choices with foldout without any previous interaction.

After the initial rating where the choices were "Bad", "Medium" or "Good", an additional row would appear with descriptive words depending on what the initial rating were. For example, if the initial rating from the user was "Bad", words such as "Disconnected" or "No sound" would appear, while if the initial rating was "Good", the words "Minor inconveniences" or "Flawless" would appear, see Figure 6.2. These alternatives would also be highlighted if the user hovered over them with the cursor. This prototype featured a "Send"-button that was made clickable after the initial rating, to enable the user to only send a Bad/Medium/Good response if they did not feel like giving any additional input.

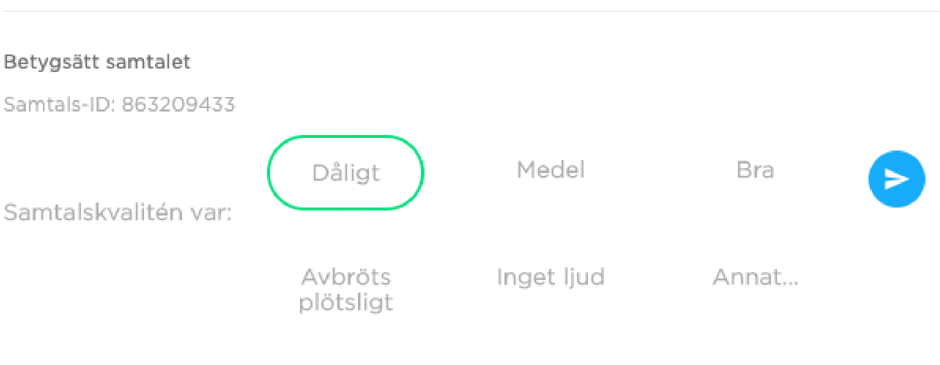

Figure 6.2: Choices with foldout after the user has rated the call "Bad".

6.4.2 Star with statistics

Star with statistics is a simple five-star system which features hover effects, meaning that stars left of the hovered star turned into a darker colour, and fading animations on rating changes. The prototype has mock-up statistics on how Telavox perceived the rating quality represented by three bars, see Figure 6.3.

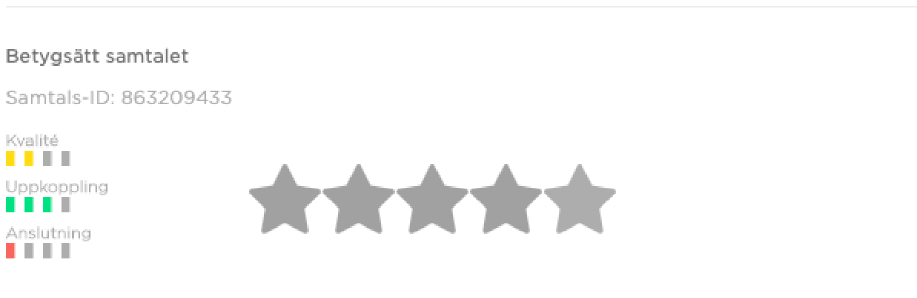

Figure 6.3: Star with statistics without any previous interaction.

The prototype features a confirming animated icon confirming and text, which can be seen in Figure 6.4. If the user was to change their rating after the initial rating, the text changed from "Sent!" to "Updated!"

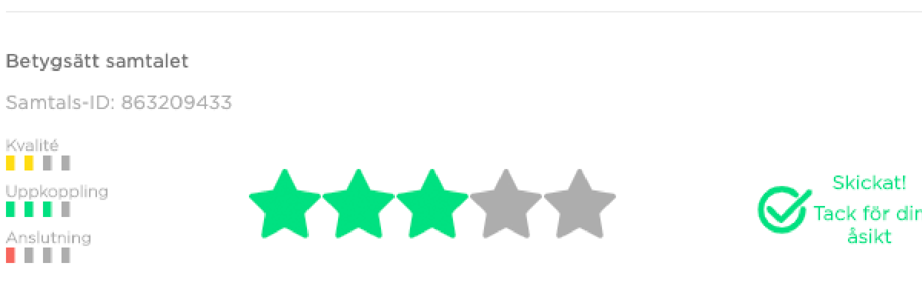

Figure 6.4: Star with statistics after the user has rated the call three out of five stars.

6.4.3 Star with statistics and foldout

Star with statistics and foldout is a combination of the two previous prototypes. If the user rated the call one or two stars, the alternatives from Choices with foldout for "Bad" showed up, three or four stars showed the "Medium", and five stars showed you the "Good". What has been added to this prototype is if additional input is given, an extra text saying "Thank you for your additional input!" shows up, see Figures 6.5 and 6.6.

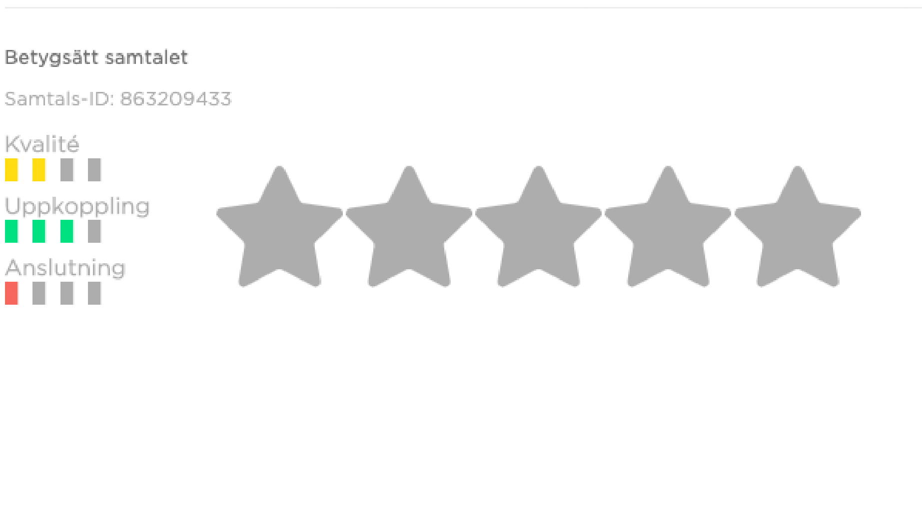

Figure 6.5: Star with statistics and foldout without any previous interaction.

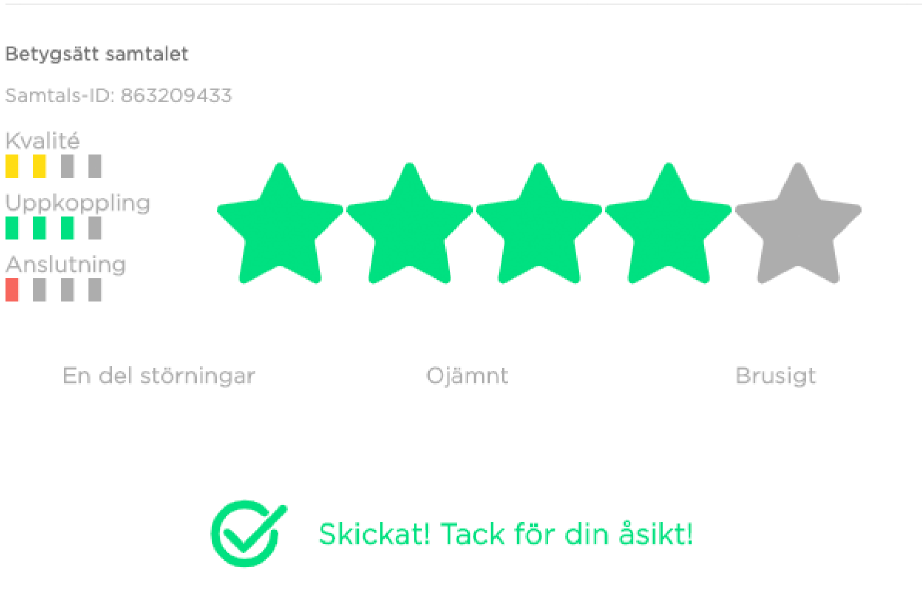

Figure 6.6: Star with statistics and foldout after the user has rated the call four out of five stars, but not given any additional input.

6.5 Execution

This time the test was executed in a more relaxed setting. We started with introducing the result of the last test, the purpose of this test and then explained that because we would take notes they would encourage thinking out loud during the test. At the end of the test we asked the test participant to rank the three systems.

6.6 Result from the Hi-Fi Phase

The result of the ranking question can be summarised with:

1. Star with statistics and foldout
2. Star with statistics
3. Choices with foldout

Table 6.1: The result of the smaller test made in Telavox's web application Flow. TP X refers to test person number X.

System	TP 1	TP 2	TP 3	TP 4	TP 5
Choices w/ foldout	3	3	3	3	3
Star w/ statistic	2	2	2	2	1
Star w/ stat. and foldout	1	1	1	1	2

Some of the comments regarding the Star with statistics and foldout scale were:

- "I like that I get the option to express myself more."
- "The star scale gives a welcoming feel. You want to use it."
- "Great to get feedback on both choices."

6.7 Conclusions from the Hi-Fi Phase

The result from the Table 6.1 was pretty united. The last prototype, Star with statistics and foldout, was chosen as the best system from all our test participants except one. And the participant not choosing it said it was because of the size of the prototype.

The size was something all of the participants commented on which lead us to making our final design smaller than the original winner. Some animation speeds were also increased for a more natural and responsive feel. All participants agreed that a Send-button was unnecessary and would rather decrease the user experience than increase it. Some comments about a free text input were also made in association with the foldout choices.

Chapter 7

Implementation Phase

7.1 Purpose

After the results and conclusions from the Hi-Fi Phase one final rating scale was chosen to be implemented. The main goal for this phase was to implement the rating scale that Telavox would then later start using.

7.2 Implementation Platform

The implementation platform was Telavox's web application Flow. We also connected our rating scale with Telavox's back-end using an API point so that the rating scores could be included in the phone call statistics.

7.3 Design Choices

Due to the comments and results from the Hi-Fi phase, the final system we decided to design was a variant of Star with statistics and foldout, but adjusted to a more appropriate smaller design, similar to the size of Star with statistics, see Figure 7.1.

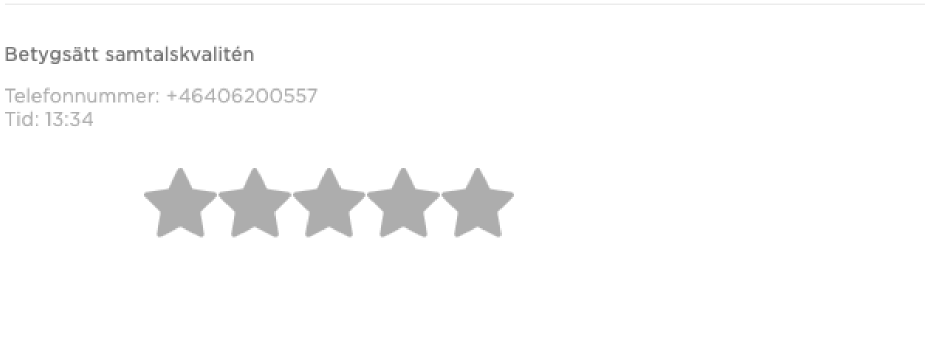

Figure 7.1: The final rating system with no previous input.

The active colour, the colour for the toggled alternatives and stars, was changed from the usual green colour that Telavox uses to the blue that Flow mostly consists of, making the rating system able to blend into the application in a better way, see Figure 7.2.

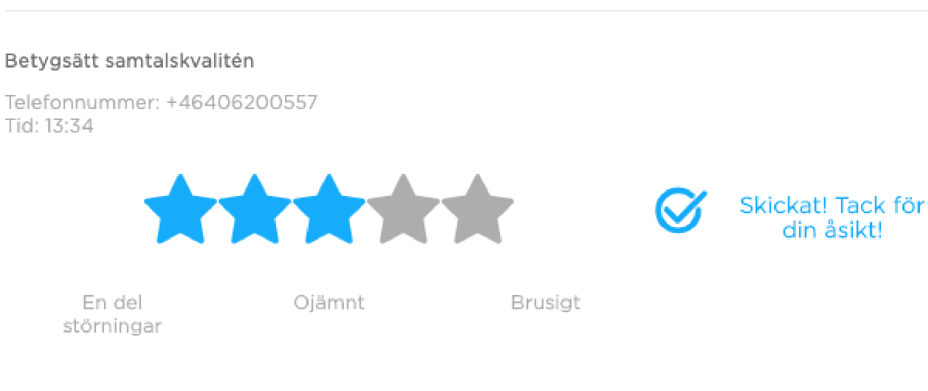

Figure 7.2: The final rating system where the user has clicked the third star. This brings up the alternatives from the "Medium" category.

The title above the rating system was also changed to "Rate the quality of the call" from "Rate the call" to further help the user understand what the rating system intends to rate. The information regarding the call was also changed from the call ID, which didn't really tell the user much information, to displaying the call's recipient and time of the call. The statistics was removed from the final implementation design, since Telavox didn't at the moment have any easy way to access the appropriate data to make the statistics meaningful.

7.4 Result from the Implementation Phase

The final rating system are shown below, in the Figures 7.3 to 7.7. The rating system also featured fade-in animations for when stars were toggled and when the text appeared. When the user updated the rating, the animation for the icon disappeared and played once again.

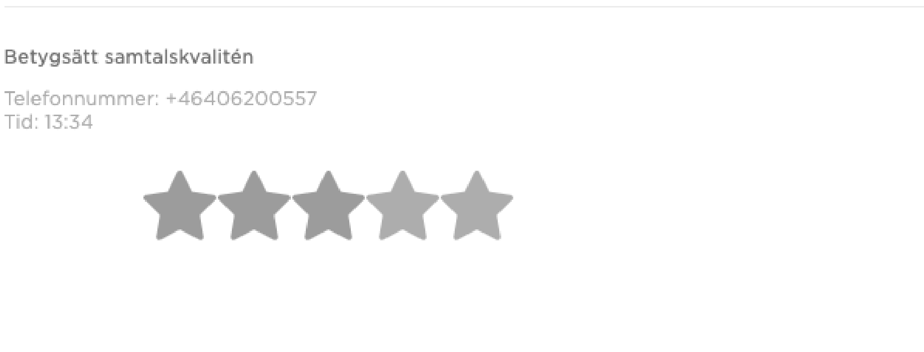

Figure 7.3: The final rating system with no previous input and the third star being hovered.

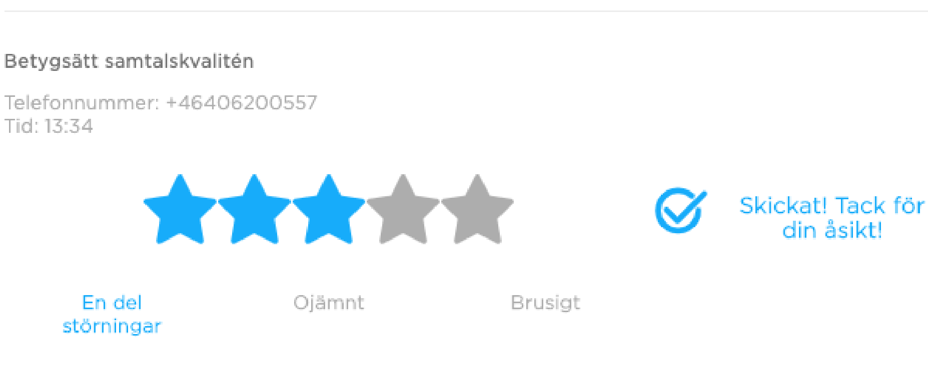

Figure 7.4: The final rating system where the user now also is hovering over the alternative "Some noise".

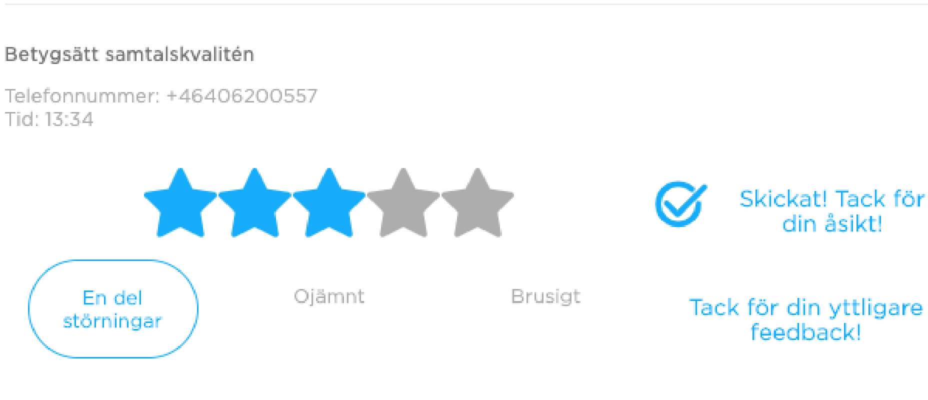

Figure 7.5: The final rating system where the user clicked the alternative "Some noise".

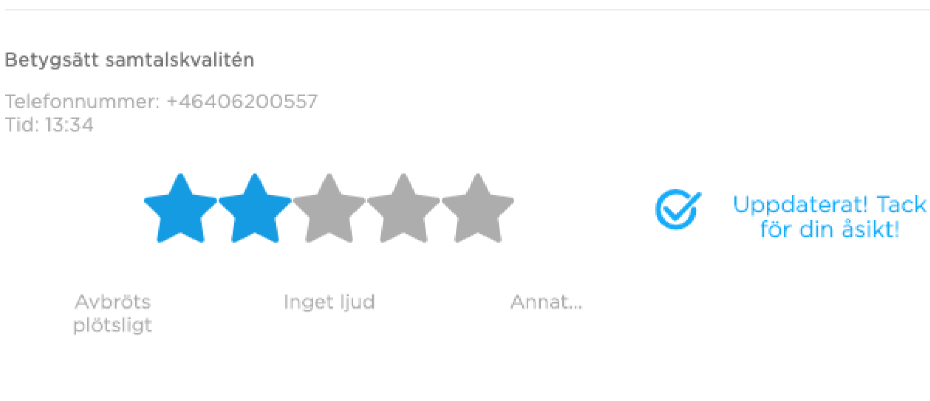

Figure 7.6: The final rating system where the user has updated their initial rating to two stars instead. This images also depicts the hover effect (darker colour the stars left of the hovered star, which is star number two from the left).

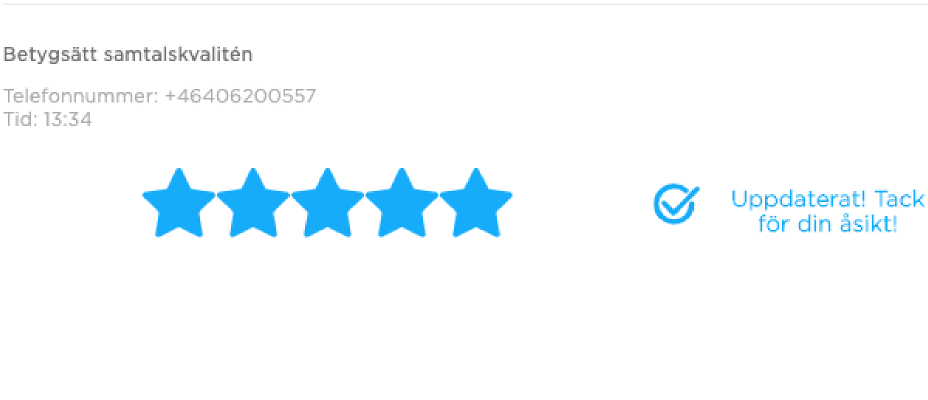

Figure 7.7: The final rating system where the user once again updates the rating, but now rated the maximum value of five stars.

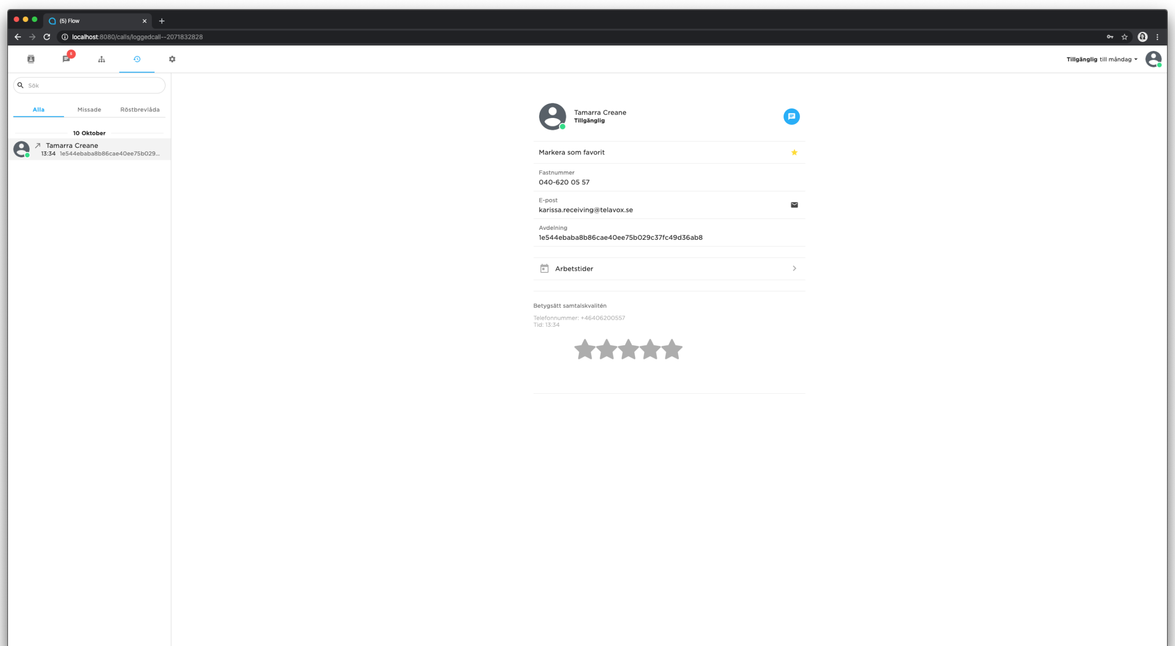

Figure 7.8: The complete view of Flow when a call history item has been clicked with our rating system.

7.5 Conclusions from the Implementation Phase

Unfortunately we did not have the time to get any result of the use frequency of the rating scale, since there was not enough time to launch the rating system into the live application.

We feel like the implemented rating scale has a very inviting design with few clicks and positive feedback as confirmation. The upside of implementing it under the call history feature is that it automatically is non-intrusive, instead it is up to the user to choose to rate the quality of the call. The rating scale is also implemented in such way that if Telavox chooses to expand the usage it would be simple to use our implementation code.

Chapter 8

Discussion & Conclusions

8.1 Research questions

8.1.1 How can different rating scales elicit different levels of user participation?

From our final questions in the Hi-Fi Phase, we got some answers that regarded the ease of use of the different systems. These pointed towards that the rating systems with the least amount of effort to use was preferred if high levels of user participation was desired. From the tests we also received a lot of comments that the rating systems with high user satisfaction was more prone to be used again.

In the Hi-Fi Phase, we also received clear indications that the systems with as few clicks as possible were preferred, meaning that e.g. the "Send"-button was not to be desired. Comments regarding ease of use for actually using of the rating system also indicated that it should be optional for additional information, since one click is what realistically some users only felt they had the desire or energy to contribute with.

8.1.2 How can different design choices elicit different levels of user satisfaction?

From the Mid-Fi Phase we found out that the most intuitive rating scale turned out to be the star scale, since it is so strongly connected to be a tool in which you rate or review things. The unary system proved to also be related closely to something that was being rated, but since the context was more vague than the usual unary rating systems (e.g. appearing under a picture on social media), its cause was lost.

We saw a strong positive trend towards the rating systems with the more experimental design of our tests compared to the baseline prototypes, suggesting that fade-in animations and hover-effects increase the intuitiveness and user satisfaction of the rating systems.

Also, if you compare the three base binary systems: Thumbs up/down, arrows up/down, and emojis happy/sad, we saw only a positive final score for the emojis. This is most likely since the emojis had a red colour representing the negative input, and the standard green for the positive, whereas the thumbs and arrows had green colour as their colour independent of input.

This argument can also be strengthened because the gap between the star scale's stars and emojis are much closer.

8.2 Discussion

The overall investigation, implementation and result were for us considered a success. From our investigation phase and Mid-Fi Phase, we learned a lot regarding how hard it can be to overcome what us humans have learned previously when encountering a familiar situation, but that acts in a different way compared to the previous experiences. We learned that we could benefit from this a lot, and therefore create combinations from regular rating systems with more functionality that still acted and worked as what the user expects, while still being in a new setting and gathering more information from the user without adding much more effort.

The rating system that was finally implemented contained a lot of improvements from ordinary star rating scale, making it more extensive and usable in the way it gathers data compared to a lot of similar systems, while still being easy to use having the option for the user to click once or twice depending on how much time or effort the user wants to spend rating the call.

8.3 End Product & Future Work

In the end we delivered our rating system to our own Telavox repository for them to later move into production, making it available for all users of Flow after a code review. Another task left to complete, which meant a substantial amount of work that we unfortunately did not have enough time to complete, was to design a proper database for them to more easily access the rating system.

When the rating system is to be moved into production, we would also advise Telavox to test the rating system against their customers because even if our test group was made as broad as possible, it's still the real customers that is the final target audience which may vary from our test group.

Something that would be in Telavox's interest is connecting a low rating score with their already existing ticket system. Making it an automatic process for the user to directly and smoothly contact Telavox's Advisors or possibly forward the issue to the development team directly, so that bad call quality and similar issues can be dealt with quicker.

The rating system that we implemented was designed in and for the desktop version of Flow, which means that for it to work in the phone based applications, modifications in how you access the rating system has to be made since the phone based applications has another kind of design when accessing the call history.

The statistic part that was tested in the Hi-Fi Phase is something that we would recommend Telavox to implement in the future, and to then test how the user reacts with the statistics presented compared to the final implementation. These statistics are also something that could be analysed with machine learning for accurate statistics calculated from not only the back-end values, but also from what users with similar back-end values has rated previously. If this algorithm was designed well, it could also be possible for tailor-made alternatives depending on the rating. Meaning that the rating system could predict if the call got disconnected or was noisy, and present that to the user. This could then be used by the algorithm to train itself from user input, thus improving itself.

After the implementation it would also be interesting to analyse the use frequency of the rating scale implemented. Which, since we didn't release our implementation live, we never got result from.

Bibliography

Abras, Chadia, Diane Maloney-Krichmar, Jenny Preece, et al. (2004). "User-centered design". In: *Bainbridge, W. Encyclopedia of Human-Computer Interaction. Thousand Oaks: Sage Publications* 37.4, pp. 445–456.

Alsudani, Farah and Matthew Casey (2009). "The effect of aesthetics on web credibility". In: *Proceedings of the 23rd British HCI Group annual conference on people and computers: Celebrating people and technology.* British Computer Society, pp. 512–519.

Bauerly, Michael and Yili Liu (2008). "Effects of symmetry and number of compositional elements on interface and design aesthetics". In: *Intl. Journal of Human–Computer Interaction* 24.3, pp. 275–287.

Bevan, Nigel (2009). "What is the difference between the purpose of usability and user experience evaluation methods". In: *Proceedings of the Workshop UXEM.* Vol. 9, pp. 1–4.

Blackmon, Marilyn Hughes, Peter G Polson, Muneo Kitajima, and Clayton Lewis (2002). "Cognitive walkthrough for the web". In: *Proceedings of the SIGCHI conference on human factors in computing systems.* ACM, pp. 463–470.

Duyne, Douglas K Van, James Landay, and Jason I Hong (2002). *The design of sites: patterns, principles, and processes for crafting a customer-centered Web experience.* Addison-Wesley Longman Publishing Co., Inc.

Gonzalez, Cleotilde (1996). "Does animation in user interfaces improve decision making?" In: *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, pp. 27–34.

Göransson, Bengt, Jan Gulliksen, and Inger Boivie (2003). "The usability design process–integrating user-centered systems design in the software development process". In: *Software Process: Improvement and Practice* 8.2, pp. 111–131.

Gould, John D and Clayton Lewis (1985). "Designing for usability: key principles and what designers think". In: *Communications of the ACM* 28.3, pp. 300–311.

Mahatody, Thomas, Mouldi Sagar, and Christophe Kolski (2010). "State of the art on the cognitive walkthrough method, its variants and evolutions". In: *Intl. Journal of Human–Computer Interaction* 26.8, pp. 741–785.

Maron, Deborah, Cliff Missen, and Jane Greenberg (2014). "Lo-Fi to Hi-Fi": A new way of conceptualizing metadata in underserved areas with the eGranary digital library". In: pp. 37–42.

Miller, Hannah Jean, Jacob Thebault-Spieker, Shuo Chang, Isaac Johnson, Loren Terveen, and Brent Hecht (2016). ""Blissfully Happy" or "Ready to Fight": Varying Interpretations of Emoji". In: *Tenth International AAAI Conference on Web and Social Media*.

Norman, Donald (2002). "Emotion & design: Attractive things work better". In: *interactions* 9.4, pp. 36–42.

O'Brien, Heather L and Elaine G Toms (2008). "What is user engagement? A conceptual framework for defining user engagement with technology". In: *Journal of the American society for Information Science and Technology* 59.6, pp. 938–955.

Page, Tom (2014). "Skeuomorphism or flat design: future directions in mobile device User Interface (UI) design education". In:

Rick, Jochen, Phyllis Francois, Bob Fields, Rowanne Fleck, Nicola Yuill, and Amanda Carr (2010). "Lo-fi prototyping to design interactive-tabletop applications for children". In: pp. 138–146.

Rubin, J. and D. Chisnell (2008). *Handbook of Usability Testing*. Wiley Publishing, Inc.

Sparling, E Isaac and Shilad Sen (2011). “Rating: how difficult is it?” In: *Proceedings of the fifth ACM conference on Recommender systems*. ACM, pp. 149–156.

Sundar, S Shyam and Sriram Kalyanaraman (2004). “Arousal, memory, and impression-formation effects of animation speed in web advertising”. In: *Journal of Advertising* 33.1, pp. 7–17.

Tourangeau, Roger and Kenneth A Rasinski (1988). “Cognitive processes underlying context effects in attitude measurement.” In: *Psychological bulletin* 103.3, p. 299.

Tversky, Amos and Daniel Kahneman (1974). “Judgment under uncertainty: Heuristics and biases”. In: *science* 185.4157, pp. 1124–1131.

Wikipedia, the free encyclopedia (2019). *Brainstorming*. URL: `https://en.wikipedia.org/wiki/Brainstorming` (visited on 12/16/2019).

Yoo, Chan Yun, Kihan Kim, Stout, and Patricia A (2004). “Assessing the effects of animation in online banner advertising: Hierarchy of effects model”. In: *Journal of interactive advertising* 4.2, pp. 49–60.

Appendix A

Diagrams and Graphs From Mid-Fi Phase Test Iteration 1.

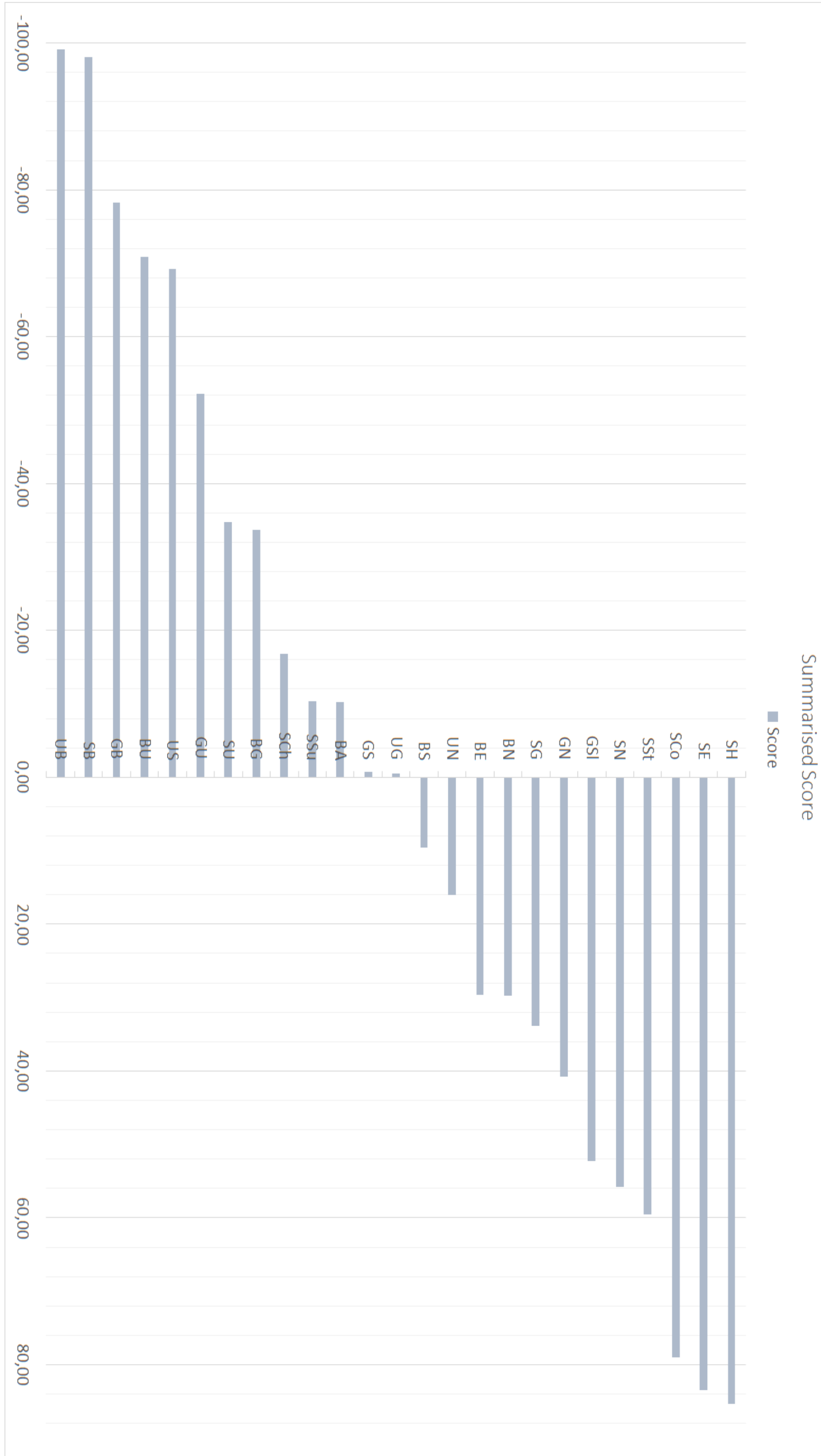

Figure A.1: Graph showing the three assessments separated for each rating system.

Figure A.2: Graph showing the three assessments separated for each rating system.

Appendix B

Diagrams and Graphs From Mid-Fi Phase Test Iteration 2.

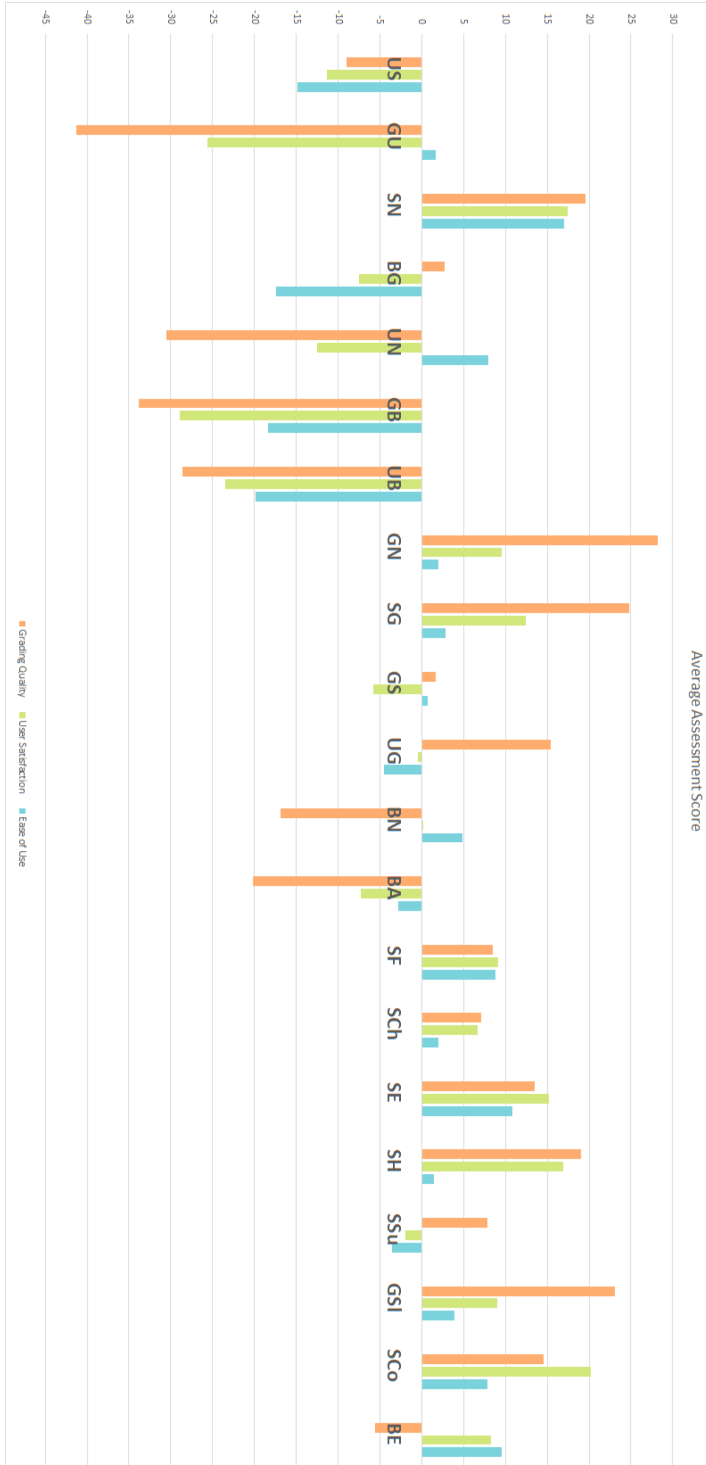

Figure B.1: Graph showing the three assessments separated for each rating system.

Figure B.2: Graph showing the summarised score for all three assessments for each rating system ordered

.

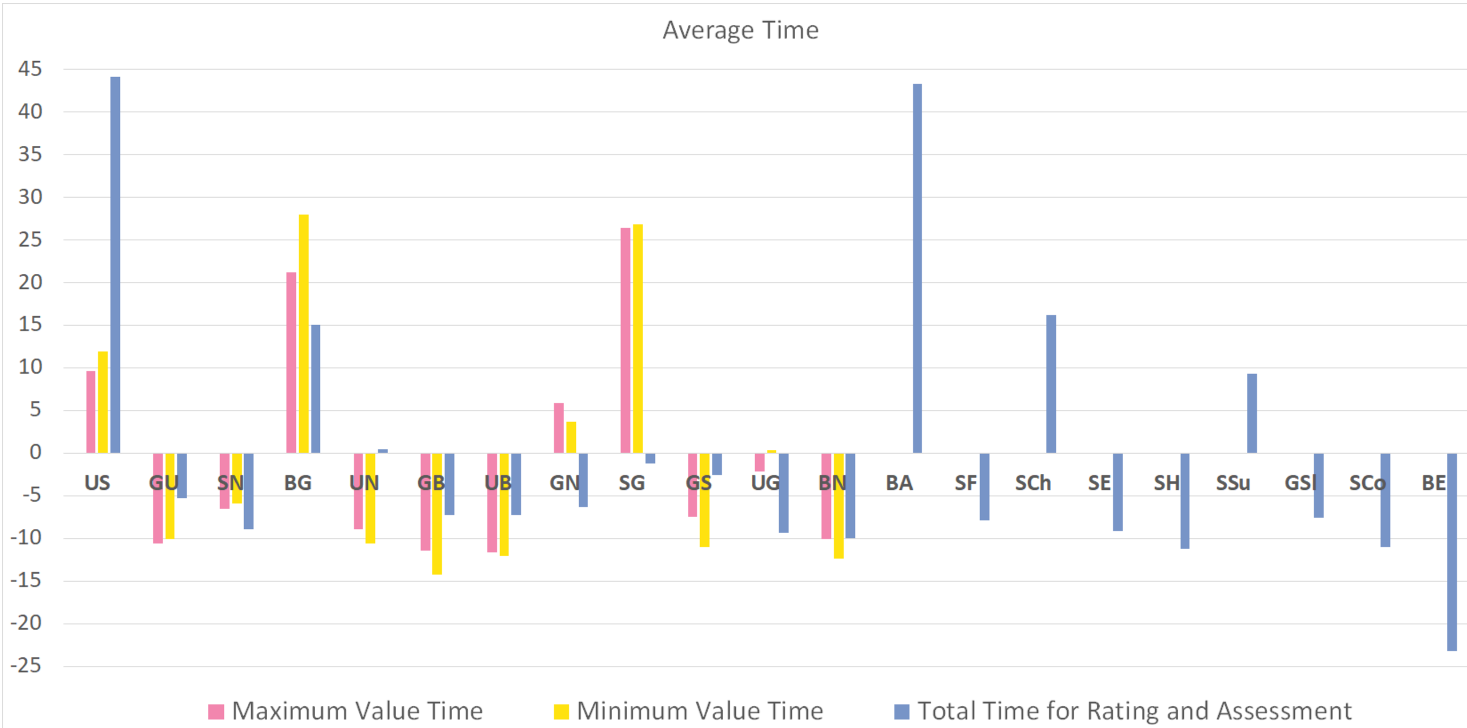

Figure B.3: Graph showing the time to achieve both the maximum and minimum goal, as well as the total time spent on the system.

Appendix C

Compilation of comments for the 21 Mid-Fi prototypes.

- US - Confusing, Rather a 10 scale than 5 and Hard to understand
- GU - Confusing, Stupid and Weird
- SN - Classic, Intuitive and Established
- BG - Non-Intuitive, A bit Unclear and Is 50 a negative value?
- UN - Boring, Do not like it and Unspecific
- GB - Weird, Unspecific and Non-Intuitive
- UB - Boring, Misses 0 as an alternative and Simple
- GN - Specific, Good and Hard to use
- SG - Specific, Non-Intuitive and Messy
- GS - Rather a 10 Scale than 5, Misses 0 as an alternative and Unspecific
- UG - Specific, Bit weird and Confusing
- BN - Unspecific, Misses 0 as an alternative and Easy
- BA - Wrong association, Unspecific and Easy to use
- SF - Intuitive, Simple and Misses the fifth star

- SCh - Different!, Misses alternative and Specific
- SE - Intuitive, Easy to Understand and Fun
- SH - Bit confusing hoover, Hard to use and Like it
- SSu - Confusing, Influencing and Unclear
- GSl - Good, Easy and Specific
- SCo - Encouraging, Nice Design and Great
- BE - Good, Like the colours! and Misses alternative

Appendix D

Compilation of comments for the three Hi-Fi prototypes.

Some of the comments regarding the Star with statistics and foldout scale were:

- "I like that I get the option to express myself more."
- "The star scale gives a welcoming feel. You want to use it."
- "Great to get feedback on both choices."

Some of the comments regarding the Star with statistics scale were:

- "Easy. Good with feedback."
- "I do not understand the calculated rating part, but I like the stars."
- "Better design than the choices. You understand the star scale instantly."

Some of the comments regarding the Choices with foldout scale were:

- "Not overly inviting, and I do not like the send button."
- "I like the different choices but they need to be better and in the right order."
- "I feel a bit boxed in with these suggestions and would greatly appreciate a free text input."