



Fraißler Gottfried, BSc

Investigation of User Behavior in Preference Acquisition

Master's Thesis

to achieve the university degree of

Master of Science

Master's degree programme: Computer Science

submitted to

Graz University of Technology

Supervisor

Univ.-Prof. Dipl.-Ing. Dr.techn. Felfernig Alexander

Co-Supervisor

Dipl.-Ing. Dr.techn. BSc. Stettinger Martin

Institute for Software Technology

Head: Univ.-Prof. Dipl.-Ing. Dr.techn. Slany Wolfgang

Graz, March 2019

This document is set in Palatino, compiled with pdfL^AT_EX₂ε and Biber.

The L^AT_EX template from Karl Voit is based on KOMA script and can be found online: <https://github.com/novoid/LaTeX-KOMA-template>

Affidavit

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master's thesis.

Graz, _____

Date

Signature

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Das in TUGRAZonline hochgeladene Textdokument ist mit der vorliegenden Diplomarbeit identisch.

Graz, am _____

Datum

Unterschrift

Abstract

The digital world is expanding continuously. We, as the consumers of the world wide web, get confronted with applications and new features every day. Nowadays, many applications use some sort of recommender system which means that the consumers get the content they want to see and get recommendations for products they like to buy. To get the information from the users, those systems use rating scales. We see them all the time while using applications on our smart phone or by browsing through the internet (e.g., rate a product bought in an online shop or like a video on a streaming platform).

This thesis examines whether the rating scale itself impacts the outcome of the rating process. Furthermore whether the scale influences our behavior when rating different products. For this, we created a user study with 138 responses from our participants. The study is split into two parts. The first part focuses on a *requirement engineering* process. Here we want see if the outcome of the process differs when using different rating scales. Half of the given scales are so called graphical scales, so we also want to find out

if one of those stands out. The second part of the study focuses on scales preferred by the participants. We ask questions about which rating scales would they use and prefer in the provided domains.

The evaluation of the first part shows that rating scales indeed have an influence on the result of a rating scenario. Also one can see that the duration participants spent on the rating process correlates (to some degree) with the usability and accessibility of the scale. The evaluation of the second part indicates that there is a rating scale that most users would like to use on any topic, but there are some aspects inside the provided domains where users clearly want to use something else. Overall, the conclusion is that one should not ignore the effect and influence of a rating scale.

Kurzfassung

Unsere digitale Welt wächst stetig jeden Tag. Wir, als die Konsumenten des Internets, werden jeden Tag mit neuen Applikationen und Funktionen konfrontiert. Viele dieser Applikationen besitzen heutzutage ein sogenanntes "Recommender System", zu Deutsch "Empfehlungssystem" als Basis. Mithilfe dieser Systeme können Dienste beispielsweise Produktempfehlungen auf online Kaufplattformen geben. Um die Information der Benutzer zu erhalten, besitzen diese Systeme Bewertungsskalen. Diese kommen uns so ziemlich in jeder modernen Anwendung unter, sei es, um ein Gespräch zu bewerten oder einem Video einem "Daumen hoch" zu geben.

In dieser Diplomarbeit untersuchen wir nun den Einfluss, den diese Skalen auf den Bewertungsprozess haben. Darüber hinaus, ob diese Skalen uns als Benutzer selbst beeinflussen. Um dies herauszufinden haben wir eine Studie mit 138 wertbaren Teilnahmen durchgeführt. Die Studie wurde in zwei Hälften aufgeteilt. Die erste fokussiert sich auf einen sogenannten "Requirement Engineering Process", zu Deutsch "Anforderungsanalyse". Hier wollen wir beobachten, ob und wie verschiedene Bewertungsschemen

Einfluss auf den Ausgang der Analyse haben. Die Hälfte der vorgegeben Bewertungsschemen sind grafische Skalen, daher soll auch herausgefunden werden, ob eine dieser Skalen besonders hervorsteicht. Die zweite Hälfte der Studie bezieht sich auf die bevorzugten Bewertungsskalen der Teilnehmer. Die Teilnehmer wählen aus verschiedenen Bewertungsskalen, für verschiedene Domänen und Attribute, ihre Favoriten.

Die Auswertung des ersten Teiles zeigt, dass unterschiedliche Skalen tatsächlich einen Einfluss auf das Ergebnis eines Bewertungsszenarios haben. Man kann auch erkennen, dass es bis zu einem gewissen Grad einen Zusammenhang zwischen der aufgewendeten Zeit um die Anforderungsanalyse durchzuführen, der Benutzerfreundlichkeit, sowie der Zugänglichkeit der Bewertungsskala gibt. Die Auswertung der zweiten Hälfte zeigt einen starken Trend auf. Es gibt eine Skala die generell bevorzugt wird, unabhängig von der Domäne und den Attributen. Es gibt allerdings Ausnahmen, die zeigen, dass auch andere Skalen für einzelne Attribute innerhalb einer Domäne bevorzugt werden.

Die generelle Erkenntnis der Arbeit ist, dass man den Effekt und den Einfluss der Bewertungsskalen nicht ignorieren sondern berücksichtigen soll.

Contents

Abstract	v
1. Introduction	1
2. Related Work	7
2.1. Preference Acquisition	7
2.2. Modelling User Preferences	10
2.3. Rating Scales	10
2.4. Survey Creation	14
2.5. Requirement Prioritization	15
2.6. Rating Bias	16
3. Study Description	19
3.1. General Structure	19
3.2. Hypotheses	32
4. Evaluation	35
4.1. Demographic Background	35

Contents

4.2. Requirement Engineering Scenario	36
4.2.1. Impact of a Rating Scale on the Outcome of the Scenario	37
4.2.2. Tendency for Graphical Rating Scales	41
4.2.3. Time Comparison of Different Rating Scales	41
4.3. Usability and Understandability of the Rating Scales	42
4.4. The Most Preferred Rating Scales	46
4.4.1. Single Domain	46
4.4.2. Multi Domain	48
5. Example User Interfaces	51
5.1. 5-Star Rating Scale	51
5.2. Thumbs Up/Down Rating Scale	53
5.3. Ranking Rating Scale	58
5.4. Heart Rating Scale	60
5.5. Slider Rating Scale	62
5.6. Categorization Rating Scale	63
5.7. When to Use Which Scale	66
6. Future Work	67
7. Summary and Conclusion	69
A. Processed User Study Data	75
A.1. Demographic Data	77
A.1.1. Gender	77
A.1.2. Age	79

Contents

A.2. Point System	80
A.2.1. Votes Converted to Points	80
A.2.2. Point Comparison	87
A.2.3. Corresponding Rank Comparison	88
A.3. Graphical Rating Scales	90
A.3.1. Votes With Stars	90
A.3.2. Votes With Hearts	92
A.3.3. Votes With Thumbs	94
A.4. Timings	96
A.5. Usability	98
A.6. Understandability	100
A.7. Single Domain	102
A.7.1. Individual Single Domain Results	104
A.7.2. Definition of Single Domains	105
A.8. Multi Domain	107
Bibliography	109

List of Figures

3.1. Welcome Screen in Every Survey	20
3.2. Demographic Questions	21
3.3. Requirement Engineering-related Questions With Ranking Scale	24
3.4. Usability-related Questions in Limesurvey	26
3.5. Understandability-related Questions in Limesurvey	27
3.6. Preferred Rating Scales-related Questions	30
3.7. Final Screen in Limesurvey	31
4.1. Gender Distribution	36
4.2. Point System for Comparison	38
4.3. Comparison Between Scaling Methods and Point System . . .	39
4.4. Comparison Between the Ranks with Point System	40
4.5. Average Time Need for Preference Acquisition in Each Rating Scale	42
4.6. Average Usability of the Rating Scales	43
4.7. Average Understandability of the Rating Scales	44

List of Figures

4.8. Preferred Rating Scales in a Single Domain Approach	46
4.9. Preferred Rating Scales in a Multi Domain Approach	48
5.1. Amazon Review Page	52
5.2. Netflix Recommendation System	55
5.3. Youtube Video Page	56
5.4. Choicla Android App	58
5.5. Jira Issue Ranking	59
5.6. Book Rating With Hearts	61
5.7. Rating With a Slider Scale	62
5.8. Categorization Scale for Game Engagement Questionnaire . .	64
5.9. Mobile Categorization Scale	65

1. Introduction

In order to get an understanding of the research problem, this chapter provides a general overview and insight on the topics of this thesis.

"Would you like to rate this app?", this is a very common question in mobile and web applications nowadays to increase the number of reviews. (Guzman and Maalej, 2014) Rating scales can be found everywhere in the modern digital world. App ratings can positively or negatively affect important aspects of how people discover apps and show how successful apps are. (Fu et al., 2013; Guerrouj, Azad, and Rigby, 2015) They not only provide feedback for the developer or service, but also reveal information about the user. Engaging the user to rate or give feedback, is one of many hard tasks in the rating and evaluation process. (Mauldin, 2014; Di Sorbo et al., 2016) Looking at it from a psychological point of view, a "like" on a social media platform, which basically is a rating of another persons picture, status update, or any other kind of shareable content, can have a big impact on the mental health of people. As discussed by Soat (2015), the rush of happiness and contentment you feel after receiving a like, is thanks to dopamine, a

1. Introduction

neurochemical known as the “reward molecule” that’s released after certain human actions or behaviors, such as exercising or setting and achieving a goal. Also, on the other side of the spectrum as the person who is supposed to like the comments and updates from others, you don’t even have to go through the physical exertion of clicking “like” to feel the rush. Often the earliest predictor of a reward, like your phone buzzing when someone posts or updates on social media, will get you a rush of dopamine.

Today’s digital services make use of the concept of *rating*. Rashid, Karypis, and J. Riedl (2005) and Xiao and Benbasat (2007) describe how the concept of rating is used for recommender systems where rating influences the items recommended for the user. Amazon¹ uses it for their item-to-item recommendation described by Linden, B. Smith, and York (2003). The whole process of collecting user preferences is called *preference acquisition* which in this case is done by item rating.

This thesis focuses not only on user behavior when rating items and the impact different rating scales have on the outcome of the acquisition process but also looks into the most liked and preferred rating scales. For this we conducted a user study which simulates a rating scenario in a requirement engineering process. We also provide a questionnaire on preferred rating scales when (1) rating items in different domains which is a single-attribute approach and (2) rating different features in one domain which is a multi-attribute approach.

¹www.amazon.com

The major results of this thesis are the following: the chosen rating scale in a requirement engineering process does influence the prioritization of the requirements, specifically, ranking-based rating scales can result in a clearer prioritization; when we look in more detail on graphical rating scales in the requirement engineering process, one can see that the average ratings differ significantly between the various graphical scales, specifically, thumbs, on average, trigger higher rating values; an interesting finding is that the rating scale with quickest average response time has the worst usability in terms of difficulty to give a rating according to the participants of the study; also an interesting finding is that understandability and usability of a scale are tightly connected; in general, people like the 5-Star rating scale the most and would prefer it in almost every rating scenario; overall, we should not underestimate the impact a rating scale can have.

1. Introduction

Outline

This thesis is organized as follows:

- Chapter 2 provides further details on related work.
It gives information about similar projects on rating scales, recommender systems, and preference acquisition.
- Chapter 3 describes how the user study was structured and provided to the test users.
For each rating scale, a survey was created. Each user had to rate in the same requirement engineering problem with one of the rating scales. After that, some questions regarding rating scales were asked on single and multi attribute domains.
- Chapter 4 focuses on the evaluation of the user study.
The user study was conducted with over 100 participants. After processing participations and filtering out invalid responses, every individual rating scale survey had at least 23 complete participations, which leads to a total number of 138 participations. The data gathered through this study is evaluated and presented.
- Chapter 5 introduces example user interfaces with the chosen rating scales.
Some tools, services, and websites are presented which are using the rating scales described in the study and an impression is given on how they are used in their services.
- Chapter 6 gives an outlook on possible future work.

It gives a short overview on future projects related to this topic.

- Finally, Chapter 7 provides a summary and conclusion.

The outcomes are shortly discussed again and a summary of the whole thesis is given.

- After the main chapters, an Appendix with the processed data is provided. The processed data that is retrieved from our survey tool is presented with tables.

2. Related Work

This chapter gives an overview of related work on the different topics related to this thesis. To get a better structuring, the topics are grouped into sections.

2.1. Preference Acquisition

Users rarely know all their preferences when they start a preference acquisition process. Ricci and Nguyen (2007) point out that users usually form their preferences during the decision making, so it is important to let them revise their preferences during the whole process. The authors explain how they tackled the problem. They designed a product recommendation methodology and implemented MobyRek, a mobile-phone recommender system used for travel products. MobyRek only supports short questions and few answer options and uses critiques, which is mentioned in Ricci and Nhat Nguyen (2005), but in their implementation it is also coupled

2. Related Work

with NutKing, a web based recommender. With this system, users can make travel plans and get recommendations on the go which can be critiqued on each recommendation cycle.

Branting and Broos (1997) introduce an automated acquisition approach of user preferences. The authors describe how a learning apprentice system is used to acquire preferences in form of preference predicates, for example, the state-preference method which can be implemented through perceptron learning. Also, two new instance-based algorithms for preference predicate acquisition are proposed. The evaluation showed that both algorithms used in a learning apprentice system to schedule astronomical observations, rapidly achieved useful levels of accuracy in predicting the astronomers preferences.

de Gemmis et al. (2009) give a general overview about preference learning in recommender systems. The paper points out the importance of recommender systems in the modern digital world and how they use algorithms for actively recommending items the user likes. Their recommendation approach comes from the idea of Information Filtering (IF). Each filtering method has it's own strength and weaknesses. A good insight on Collaborative Filtering (CF), Content-based Recommenders (CB), Knowledge-based Recommenders (KB) and hybrid systems is also given. The authors explain techniques for learning user profiles. These techniques can be split into two classes like one can split recommendation techniques into model-based and memory/heuristic based. The classes for learning user profiles are *offline* and *online* learning. Offline methods are better in systems where user

2.1. Preference Acquisition

preferences change slowly. Online methods are the more common ones and used for real-time recommendations.

User biases can influence the outcome of recommender systems. Freyne, Berkovsky, and G. Smith (2013) examine the characteristics of a data set consisting of 100,000 ratings where users were rating on a collection of recipes. The data set reflects a stable user bias towards certain features of the recipes (cuisine type, key ingredient, and complexity), which means that people would rate a recipe higher as soon as they see a certain ingredient. Knowing that a bias exists, the authors exploit this knowledge. They design and evaluate a personalized rating acquisition tool based on active learning. This helps dealing with user biases, creating high-value information, and reduces prediction errors with new users.

2. Related Work

2.2. Modelling User Preferences

In a growing digital world it is important for users to quickly get the information they need. Currently we are moving to a more personalized internet experience, thus recommender systems are used. Bollen (2015) approaches the problem of good recommendations with presenting an interaction model of preference construction. This takes into account the interaction between recommender systems and user characteristics together with contextual properties.

Dastani et al. (2005) brings the question, how does one model user preferences, to electronic commerce (e-commerce). To answer that question, the author introduces a generic mediating agent architecture. A suggestion is made that the preference of e-commerce participants can be modelled by learning from their behavior. Inductive logic programming (ILP), a machine learning method, is used by mediating agents that generates a hypothesis and takes logical theories as input. With this method, it is possible to detect regularities in the behavior of people and automatically induce a hypothesis about their preferences.

2.3. Rating Scales

How you design a survey or a form will affect the answers you get. DeCastellarnau (2018) gives a good overview of different survey response scales and

2.3. Rating Scales

which one to use. Generally, different types of data require different types of scales. The author classifies the different scales into different models, for example, "Dichotomous", "Rating Scales" and "Semantic Differential Scales". *Dichotomous scales* only have two choices. This could be, for example, "Yes or No" and "True or False". There is good value in not allowing a neutral option in long surveys. *Rating scales* are the most familiar one. The most common scales are "1-10", "1-7" or "1-5" scales where 1-5 represents the Likert scale. In Likert scales, the highest value needs to be the most positive one. The author also mentions that there is more variance in larger scales so the outcome using different scales changes. *Semantic differential* scales consist of an interval scale with a dichotomous word on the end of both spectrums (e.g. "Inexpensive and Expensive" with a neutral point in the middle of the scale). They measure a more specific attitudinal response.

An interesting question about rating scales is answered by Keusch and Yan (2015). An experiment was made to show if the direction of a rating scale has an impact on the result. The direction of a rating scale indicates if a rating scale starts with the most negative or most positive value. The experiment used a zero to ten scale for rating different countries. The result shows an impact. Countries with a higher rating in general received higher rating on average when stating with the most positive value. This appears because due to respondents' use of anchor and adjustment heuristics.

An experiment by Funke, Reips, and R. Thomas (2010) shows that slider rating scales are for smart people. In the experiment, participants had to complete a survey on health-related products. The participants had to use

2. Related Work

one of four different rating scales. Either a horizontal slider, a vertical slider, horizontal radio buttons, or vertical radio buttons. Each scale represented a 7-point scale. The result shows that the break-off rate and the response time are significantly higher when people use one of the slider scales. Problems with sliders are prevalent in participants with less than average education. This indicates that sliders are more challenging to use. Another outcome is that when using radio buttons, participants tend to choose the middle category more often than when using a slider. This is probably the case because the slider was already in the middle as default value and participants felt like they have to change the value and move the slider.

M. Thomas and Kyung (2018) takes a closer look on how the response format in payments influences our willingness to spend more money on a product. For this, a user study with several bidding scenarios (e.g., bidding on eBay¹) was created where half of the participants used a text box and half of them used a slider to input the bid. The results show that in ascending payment formats, people give higher bids on the products when using a slider than people using a text box where they input the bid manually. Also in decreasing payment formats, people give lower payments when using a slider than people using a text box. This is due to the end *point assimilation effect* which means that payments elicited on slider scales tend to be assimilated toward the end point of the response range.

Different rating scales lead to different user preferences when evaluating

¹www.ebay.com

2.3. Rating Scales

items in recommender systems. Recommender systems should be able to deal with ratings from different scales when the opportunity is given to choose a preferred scale. Gena et al. (2011) presents experiments regarding the impact of rating scales on user behaviour. In this context, mathematical normalization is not enough when mapping different rating scales together. Participants in the study had to rate the same recipe with different scales. Afterwards, the result of each scale was converted into a normalized zero to one scale. They calculated a coefficient out of their results which represents the ratio between the average ratings of each scale. Although this ratio may depart considerably from mathematical proportion, the general outcome was that one should always have in mind that different rating scales can affect the outcome. The authors mention some unexpected outcome with the user ratings. Users rated the same items differently with the same rating scale under different treatment conditions. The authors point out that their rating scenario might not have been realistic enough because of the re-ratings and this can be problematic for the final result. That is why we will also focus on this topic in this thesis.

Now after collecting a lot of data with rating scales, one would need methods to calculate the total scores. A blog entry, on the website of the survey tool "Cognito Forms²", gives an insight on three easy methods to calculate scores in a one to five or any equivalent form of a one to five scale (e.g., from very poor/unsatisfied to very good/satisfied). The first and easiest method is a total score where one just sums up all the points of the questions. In the

²<https://cognitoforms.com>

2. Related Work

presented article, all points of all questions are summed up. The second method is a weighted score. This is an extension of the first method. The total score is divided by the number of questions and can also be combined with multiple rating scales to get the average number for the questions over all rating scales. One can also weight the different scales by a parameter. The third method is percentages, which is an alternative way to display the total scores. The authors also argue that the method which fits the needs the most should be chosen and that more alternatives are needed. The outcome is that we need more research on more diverse response methods.

2.4. Survey Creation

Drosos, Tsotsolas, and Manolitzas (2011) talk about the Customer Satisfaction (CSAT) metric which is the most commonly used metric for designing a survey for a market research. CSAT can use various different scales, but the most commonly ones are Likert scales where the lowest number resembles very dissatisfied and the highest number resembles very satisfied. The respondents are classified into three groups, "dissatisfied", "neutral" and "satisfied". As advantages, one can mention the easy way to implement it into your survey. CSAT is very flexible. It can be used for overall experience or specific category ratings. Also, CSAT allows to ask follow up questions to get a better understanding of what the customer wants.

Nowadays, researchers should think more carefully about the response

2.5. Requirement Prioritization

format while creating a new questionnaire. According to (Wetzel and Greiff, 2018), it appears that the step of choosing the response format gets little to no attention and often gets rushed. Most of the time, constructors rely on scales that worked well in the past, even if the chosen scale doesn't fit the needs. An example is that constructors tend to implement a "strongly agree/disagree" scale as soon as psychological questions appear even if this type of scale is not convenient for other questions.

2.5. Requirement Prioritization

An important step in the whole requirement engineering process is *requirement prioritization*. Hasan et al. (2010) remarks that the outcome of the prioritization largely depends on the prioritization method. The authors talk about different techniques, namely Analytic Hierarchy Process (AHP), Hierarchy AHP, Minimal Spanning Tree, Bubble Sort, Binary Search Tree (BST), Priority Group, Planning Game (PG), 100 points method and Planning Game combined with AHP (PGcAHP). To find the best one, those techniques were tested in an experiment. With different criteria in mind (e.g., easy to use, certainty, accuracy) the result showed that out of the mentioned techniques, PG is the optimal one for prioritizing requirements. The PG has a good scalability and provides accurate results because you only look at one requirement and its complexity at a time and not compare it to the other requirements which means the time to prioritize n requirements is n comparisons. Pinna et al. (2003) describes the usual implementation of PG.

2. Related Work

The user writes a story on an index card and declares its prioritization. The developer then rates the complexity of the story with story points. The user then decides in each iteration which stories are going to be implemented. After each iteration, the PG is repeated.

2.6. Rating Bias

Online reviews are very helpful when it comes to decision making. But these reviews are a double-edged sword. According to Askalidis, Kim, and Malthouse (2017), the reviews help consumers to make a more informed decision, but over-represent the most extreme views. This means that most people only review a product when they really like or really hate it. Thus, moderate views are almost not represented. This is a common problem in most online reviews. The authors also describe an experiment where the goal is to find a way how to avoid the extremely biased reviews and get people with moderate views to review. In the study, participants were asked to do reviews about their employer on a job site. Some of the participants received money for the reviews, some got a motivational message which said that they help other job seekers. The outcome is that people are more likely to review when they're reminded that doing so helps other job seekers. Also, pro-social incentives led the distribution of reviews to be less biased, creating a more normal bell-curve distribution of reviews because more people give moderate reviews.

2.6. Rating Bias

Rating bias is especially a problem in employee performance measuring. Most of the time, we ask supervisors, subordinates and peers who work with the employee to determine his or her abilities. According to Holzbach (1978) and O. Kingstrom and E. Mainstone (1985), we are affected by a variety of rating biases when we make our ratings which makes it hard to determine the true performance of the employee. The author also describes different types of rater biases. The halo effect occurs when the rated employee has one single attribute that stands out in a positive way, the overall rating will be higher. The opposite of the halo effect is the horn effect which means, there is one attribute that stands out very negatively. The whole rating will be dragged down. The central tendency bias is an effect where the rater tends to always pick the middle on a rating scale. Another effect is the leniency bias. This means that one goes "too easy" on the employee so that all scores will be very high. The opposite of the leniency bias is the strictness bias where one goes "too hard" on the employee. The recency bias occurs when a recent event influences the rating. Maybe the employee had a good and productive week, so one might tend to overrate or underrate if the employee had a bad week. The last bias is the similar-to-me effect. This means that people rate others higher if they are similar. Men rate men higher and women rate women higher. All kind of similarities are influencing the rating.

In contrast to the discussed work, the focus of our work is the following: investigating the rating behavior of the participants in our study, to find out if different rating scales influence our rating behavior and hence the

2. Related Work

outcome of a requirement prioritization process; evaluating the usability and understandability of different ratings scales to see if there is correlation between these two factors; taking a deeper look into graphical rating scales, to find out if different icons (stars, hearts, thumbs) impact the average ratings, and evaluating the preferred rating scales for different domains, to see if an overall most preferred rating scale exists.

3. Study Description

This chapter gives insights of how the user study was designed, which questions were asked and what the goals of each part of the study were.

3.1. General Structure

The user study has been conducted with the survey tool Limesurvey¹, which provides the relevant features to conduct a user study in this application field. The survey is organized into six different parts. First, the participant sees a welcome screen as shown in Figure 3.1.

¹www.limesurvey.org

3. Study Description

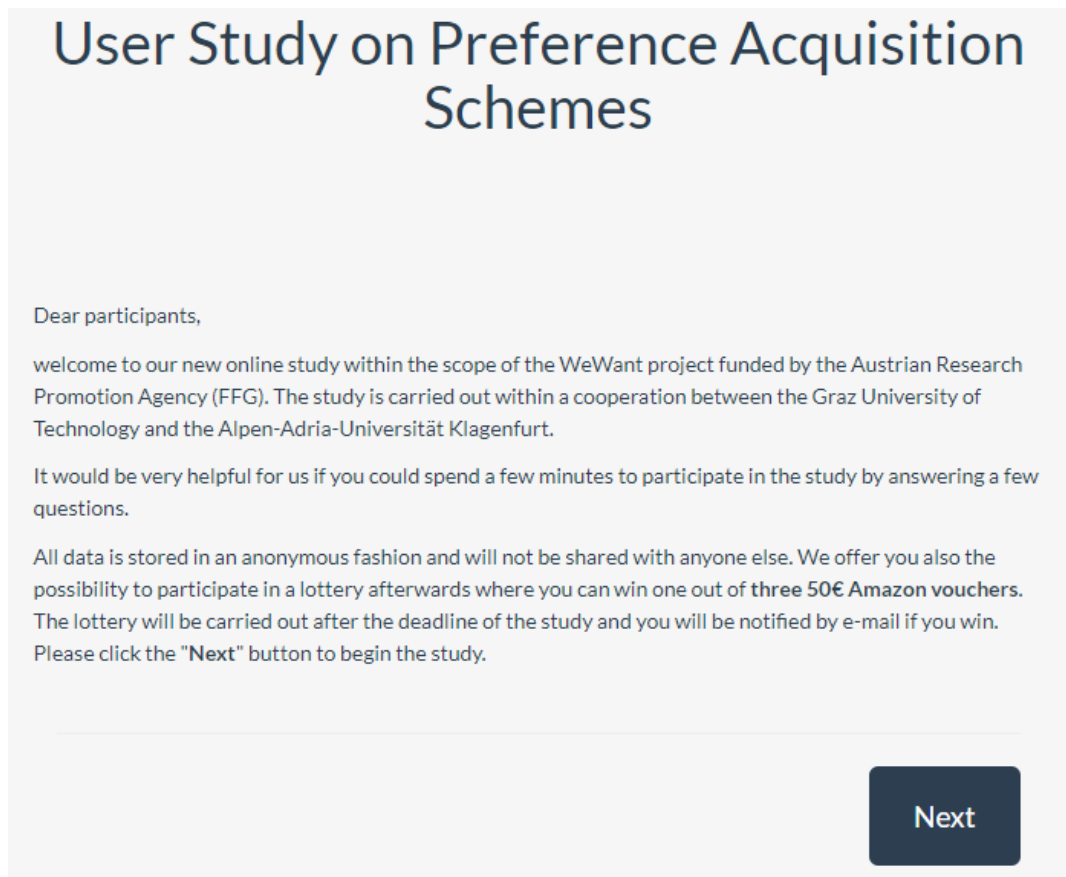


Figure 3.1.: The welcome screen every participant sees at the start of the survey

Next, we ask the participant demographic questions. This is more or less standard in every survey. Also, it helps the user to get into the survey with an easy to answer question. For the study itself it is important to ask these questions because they enable a differentiation between different sub-groups. (L. Hughes, Camden, and Yangchen, 2016) Those questions are:

3.1. General Structure

- "Gender?"
- "Age?"

In Figure 3.2, one can see the questions regarding age and gender in Limesurvey.

The following data is only used as background information for this user study and will not be shared.

* Gender?

Female Male

* Age?

Next

Figure 3.2.: The demographic questions screen in Limesurvey

3. Study Description

After this, the participant comes to the first real task which has the following description:

"Assume you are participating in a project that has the goal to develop an ONLINE COURSE REGISTRATION system for a university. A list of requirements (requirement A, B, C, D, E, F, G, H, I, J) is determined to implement the system. Please prioritize these requirements regarding their importance from your point of view."

In this scenario, each participant had to rate with one random chosen rating scale. Those rating scales were:

- 5-Star (1 to 5)
- Thumbs (5 thumbs from thumb down to thumb up)
- Ranking (Drag and drop)
- Hearts (1 to 5)
- Slider (1 to 100)
- Categorization (5 different groups from very low to very high importance)

This means that there were six different (in regards to the rating scale) surveys and every participant gets one randomly assigned when clicking the provided URL (a generic URL which has a link distributor build in and redirects to the different surveys). In Chapter 5, one can see example user interfaces of the chosen rating scales.

3.1. General Structure

Independently from the given scale, every participant had to rate the same requirements in the scenario. Those requirements are:

- *Requirement A*: Students should be able to search for the detailed course information, such as course content, semester, and academic year
- *Requirement B*: Students should be able to view course materials
- *Requirement C*: Students should be able to register for a course
- *Requirement D*: Students should be able to read information about the lecturer of the course
- *Requirement E*: Students should be able to print out the certificate of course assessment
- *Requirement F*: Students should be able to see a list of their registered courses
- *Requirement G*: Students should be able to see information of equivalent courses
- *Requirement H*: Students should be able to check the statistical evaluation of exam results
- *Requirement I*: Students should be able to take a look at the course evaluation
- *Requirement J*: Students should be able to send e-mails to all other students who are taking the same course

The questions were provided in a random order. In Figure 3.3, one can see how the questions were provided with the ranking rating scale.

3. Study Description

Please rank the requirements in the descending order of the importance, i.e., *the requirement at the top of the list has the highest importance.*

Double-click or drag-and-drop items in the left list to move them to the right - your highest ranking item should be on the top right, moving through to your lowest ranking item.

! This question is mandatory

! Please rank all items.

Your choices	Your ranking
Requirement F: Students should be able to see a list of their registered courses	Requirement J: Students should be able to send e-mails to all other students who are taking the same course
Requirement G: Students should be able to see information of equivalent courses	Requirement H: Students should be able to check the statistical evaluation of exam results
Requirement I: Students should be able to take a look at the course evaluation	Requirement E: Students should be able to print out the certificate of course assessment
Requirement C: Students should be able to register for a course	Requirement A: Students should be able to search for the detailed course information, such as course content, semester, and academic year
Requirement B: Students should be able to view course materials	
Requirement D: Students should be able to read information about the lecturer of the course	

Figure 3.3.: The requirement engineering questions with the ranking rating scale

It is important to know that *Limesurvey* can track the time people spent on the individual pages in a survey. This is very important to the study because knowing how long it takes participants to rate the scenario with the different scales on average is helpful when evaluating the results. Also,

3.1. General Structure

at this early phase of the survey, the participants did not know that the outcome of the rating is not as important as the user behavior with the different scales itself, so they might rate more naturally.

Next, there were some questions about the convenience of the used scale which the user had to answer. There were also fields where the user can explain his/her choices. The question structure looked like this:

- How do you assess the USABILITY of the rating scale? (1: very difficult, 5: very easy)
 - Please provide the explanation of your assessment regarding the USABILITY of the rating scale
- How do you assess the UNDERSTANDABILITY of the rating scale? (1: completely not understandable, 5: completely understandable)
 - Please provide the explanation of your assessment regarding the UNDERSTANDABILITY of the rating scale

In Figure 3.4 and 3.5, one can see how the convenient questions were provided in Limesurvey.

3. Study Description

Please answer the following questions:

How do you assess the **USABILITY** of the rating scale? (1: very difficult, 5: very easy)

! Choose one of the following answers

Please choose... ▼

Please provide the explanation of your assessment regarding the **USABILITY** of the rating scale:

Figure 3.4.: The questions about usability provided in Limesurvey

3.1. General Structure

The image shows a Limesurvey question interface. It consists of three main sections: a dark blue header with the question text, a white dropdown menu, a dark blue header with a sub-question, and a white text input area. At the bottom right, there is a dark blue 'Next' button.

*** How do you assess the UNDERSTANDABILITY of the rating scale? (1: completely not understandable, 5: completely understandable)**

Choose one of the following answers

Please choose...

Please provide the explanation of your assessment regarding the UNDERSTANDABILITY of the rating scale:

Next

Figure 3.5.: The questions about understandability provided in Limesurvey

3. Study Description

The goal of the next question in the survey was to better understand which rating scales are the most liked ones in general. For this we used a single domain mixed with a multi domain approach. Two out of three participants had to answer single domain questions. A single domain question focuses on one specific domain (e.g., computers) and asks about the preferred scale for different attributes of that domain (e.g., processor power, graphic card, and memory for computers). People are used to have the same rating scale for different attributes, so we decided to have more single domain questions than multi domain questions to gather more data on single domain research. That means the majority had to answer three questions about which rating scale they would use if they had to rate a specific attribute in one domain. The single domain questions were: "Assume you want to rate [an attribute]. According to your opinion, which rating scale(s) is/are optimal to articulate your preference regarding [an attribute]". The attributes were:

- "Expertise", "Working Attitude" and "Achievements" of an employee
- "Effective Resolution", "Weight" and "Price" of a camera
- "Accessibility", "Landscapes" and "Weather" of a destination
- "Location", "Price" and "Construction Quality" of a flat

In detail this means that every participant had to give their preferences on each of the three attributes of one of the domains.

The other third of the participants had to answer the multi domain questions which means that they received a wider range of domains, but only one attribute in each domain where they had to vote for their preferred rating

3.1. General Structure

scale. The multi domain questions were: "Assume you want to rate [an attribute]. According to your opinion, which rating scale(s) is/are optimal to articulate your preference regarding [an attribute]". The attributes were:

- "Songs played at a fitness center", "Movies you watched" and "Games you played"
- "Logos in a logo design competition", "Photos shown on social network websites" and "Papers presented in a conference to choose the best paper"

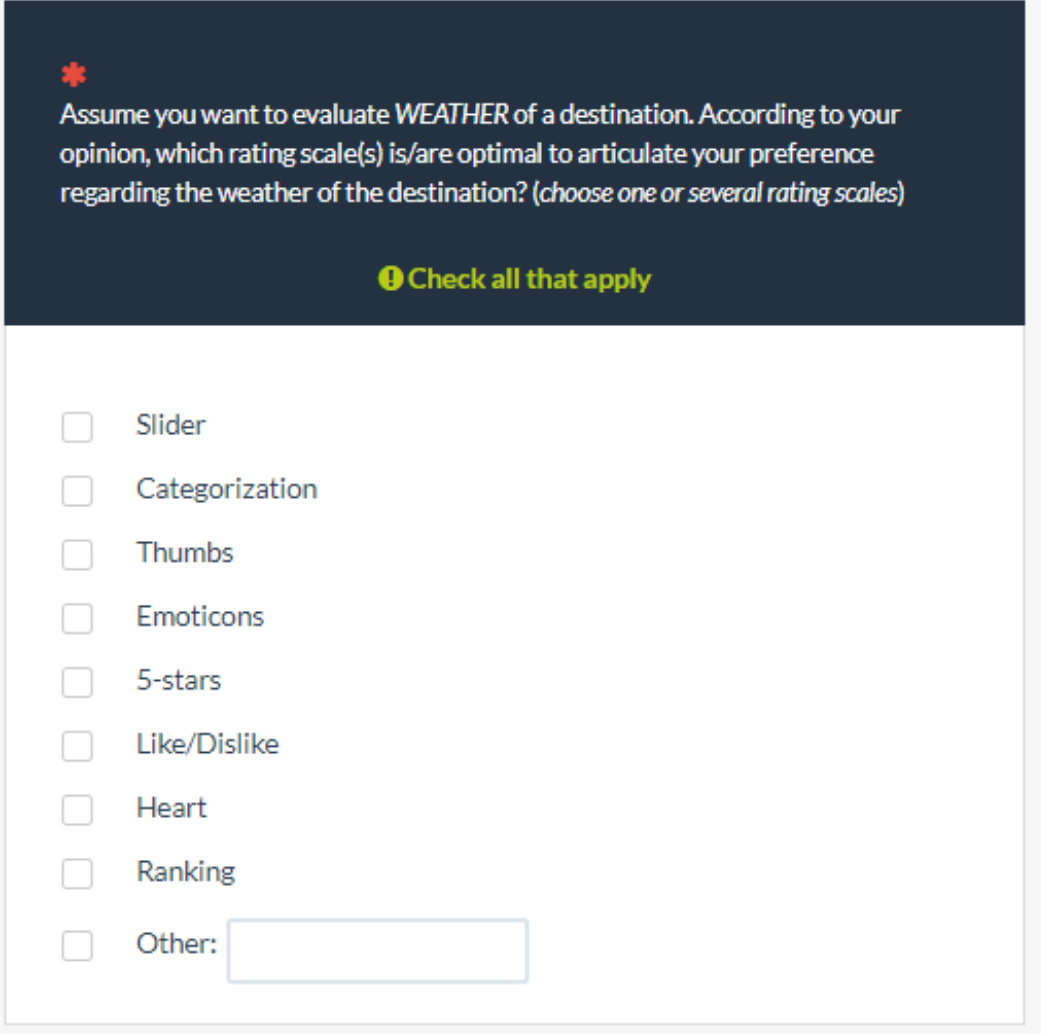
In detail this means that each participant had to give their preferences on one attribute of three different domains.

Each participant had to choose between the following rating scales in both, single and multi domain questions (Note: There were more scales than used in the requirement engineering scenario):

- 5-Star: A rating from 1 to 5 using stars
- Thumbs: A rating from 1 to 5 using thumb icons
- Hearts: A rating from 1 to 5 using hearts
- Slider: A rating from 1 to 100 using a slider
- Like/Dislike: A rating with like or dislike option
- Emoticons: A rating from 1 to 5 using emoticons
- Ranking: Ordering the items via drag and drop
- Categorization: Categorizing the items into different groups
- Other (With a field to type it in)

3. Study Description

In Figure 3.6, one can see how the question on the preferred scale looked like in Limesurvey.



The image shows a Limesurvey question interface. At the top, there is a dark blue header with a red asterisk icon and the following text: "Assume you want to evaluate WEATHER of a destination. According to your opinion, which rating scale(s) is/are optimal to articulate your preference regarding the weather of the destination? (choose one or several rating scales)". Below this header, there is a yellow button with a question mark icon and the text "Check all that apply". The main content area is white and contains a list of nine options, each with a checkbox: "Slider", "Categorization", "Thumbs", "Emoticons", "5-stars", "Like/Dislike", "Heart", "Ranking", and "Other:". The "Other:" option is followed by a text input field.

Figure 3.6.: Question on the most preferred rating scales

The last part of the survey is an appreciation message and an optional e-mail

3.1. General Structure

field where the participant can enter his e-mail address to join in a raffle. Users who were willing to join the raffle could win one of three Amazon vouchers. The participants also had to press a "submit" button to finally submit the answers. Also due to Limesurvey, participants had the option to clear their answers at any point and restart from scratch. In Figure 3.7, one can see the final screen of the survey.

Thank you for your participation. Now you can win one out of three 50 Euros Amazon vouchers. If you want to participate, please provide your e-mail address in the following text box. Your e-mail address will only be used for the purpose of the lottery. Then click "Submit" to complete the study.

Your E-Mail address (optional):

Submit

Figure 3.7.: The last screen of the survey in Limesurvey

3. Study Description

3.2. Hypotheses

With the created user study, we wanted to answer several questions. For this, we created hypotheses which will be analyzed in Chapter 4. These hypotheses are:

- *Hypothesis 1*: The chosen rating scale in a requirement engineering process has an influence on the result in terms of the prioritization of requirements.

Different rating scales have different granularity. With this hypothesis we wanted to evaluate, how big the influence of a scale really is.

- *Hypothesis 2*: Regarding graphical rating scales, the chosen icon will influence the average rating.

Different systems using different icons for their graphical rating scales. With this hypothesis we wanted to evaluate, if a certain icon drags the average rating of requirements up or down.

- *Hypothesis 3*: There is a correlation between usability and understandability of a rating scale and also with the time a user needs to complete a survey on the basis of a specific rating scale.

People might have a hard time using a certain rating scale. If a scale is too hard to understand and/or to use, it might have an impact on the time one needs to rate an item. Participants will quickly click through surveys or will never finish. With this hypothesis we wanted to evaluate, if understandability correlates with the usability of a scale and how these two factors influence the completion time.

3.2. Hypotheses

- *Hypothesis 4*: There exists a certain rating scale which is overall the most preferred rating scale, independently from the given domain. The 5-Stars rating scale is the most common used rating scale nowadays, we don't know if it is also the most preferred. With this hypothesis we wanted to evaluate, if a certain scale stands out in terms of preference.

4. Evaluation

This chapter provides the evaluation results of the user study. At the beginning, it gives a small overview on the demographic data. Thereafter it takes a look on the requirement engineering scenario as well as the usability and understandability questions. At the end, it breaks down the single and multi-attribute questions and lists the rating scales most preferred by the participants.

4.1. Demographic Background

For the evaluation we only took fully completed participations into account. Because the study was split into six different surveys (one survey per rating scale in the requirement engineering scenario), we had to balance out the number of participations. At the end, we were able to take 23 participations of each survey into account for the evaluation. This leads to a total number of 138 entries in the study. The average participant age was 23,92 (SD: 3,47)

4. Evaluation

where the oldest participant was 38 and the youngest participant was 14 years old. Gender wise, one can see a male dominance in the participants with 112 out of 138 being male which is 81%. A graphical percentage representation of the gender distribution can be seen in Figure 4.1.

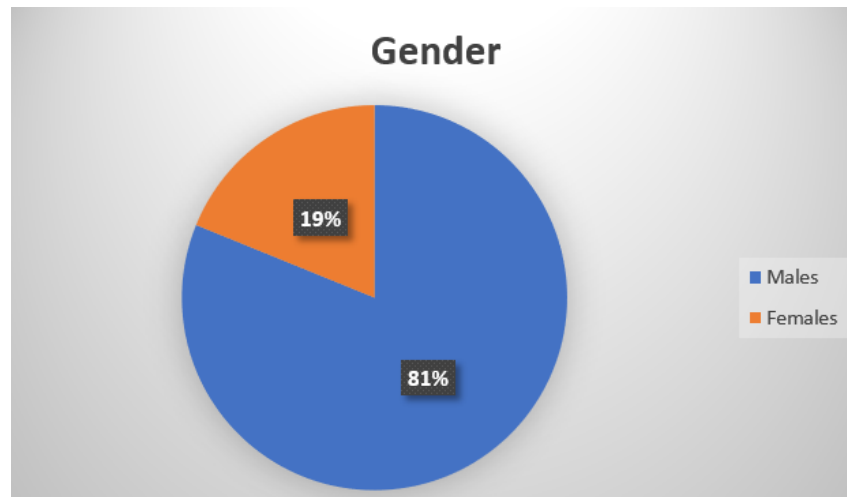


Figure 4.1.: Graphical representation of the gender distribution in the study

4.2. Requirement Engineering Scenario

For the requirement engineering scenario we wanted to evaluate several things. First, one needs to evaluate if the used rating scale has an impact on the prioritization of the requirements in our scenario. Second, one needs to evaluate if people tend to give higher or lower ratings when using different graphical rating scales (Stars, Hearts, Thumbs). Finally, we want analyze the time effort participants spent while using a certain rating scale to see if this

4.2. Requirement Engineering Scenario

effort correlates with the usability and/or understandability evaluations of the scale.

4.2.1. Impact of a Rating Scale on the Outcome of the Scenario

The first big question to answer is, if the chosen rating scale has an impact on the outcome of the rating. First we need to find a way to compare the different rating scales used in the surveys. For this, a competitive point system as it is used in sports and other competition is introduced. For every participant, the order of ranked requirements is defined, e.g., from the top rated requirement to the lowest rated one. The top requirement gets ten points, the lowest rated gets one. If two requirements have the same rating, both get the same amount of points and the next rank is skipped. For example, two requirements are top rated. Both have a rating of five stars. Those requirements take the first place and receive 10 points. The next best rated requirement receives eight points and is third placed in the order. After this, the average points of all participants for each requirement is taken and the final order of requirements (the importance) is defined by the the average values from top to bottom. This can be seen in Figure 4.2 where the point system is applied to the ratings retrieved with the 5-Star rating scale.

4. Evaluation

	Points for each requirement per user																				Sum	Avg	Rank				
A	5	10	10	10	10	10	10	10	7	4	6	10	5	10	3	10	6	10	10	8	10	7	10	7	188	8,17	4
B	10	6	10	10	10	10	8	10	10	4	10	10	10	10	7	6	10	10	8	10	10	10	7	206	8,96	3	
C	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	230	10	1	
D	1	10	2	10	4	6	8	3	4	10	2	3	3	10	7	4	5	10	3	10	1	5	5	126	5,48	7	
E	10	4	6	5	5	4	6	4	6	10	6	10	6	3	5	1	5	10	3	10	7	10	10	146	6,35	5	
F	10	10	10	10	10	10	4	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	224	9,74	2	
G	10	4	6	4	10	2	6	7	4	4	6	2	6	6	5	4	2	10	6	10	4	5	5	128	5,57	6	
H	4	4	4	4	4	2	1	2	10	4	6	10	6	3	5	10	10	10	3	10	4	5	5	126	5,48	7	
I	4	6	2	4	4	6	4	7	6	4	6	5	2	6	2	10	5	10	6	10	4	2	5	120	5,22	9	
J	4	4	4	4	1	4	4	2	1	6	2	1	3	6	1	4	1	10	6	10	7	1	1	87	3,78	10	

Figure 4.2.: The point system applied on the 5-Star rating scale

Each column represents the rating of a single user. "Sum" is the amount of points a single requirement gets from all participants using the same scale. "Avg" represents the average amount of points for a requirement. "Rank" simply represents the corresponding rank of the requirement.

This method ensures that all votes from the participants are taken into account when determining the winning requirement and the importance order. With every rating scale transformed into the point system, the comparison between them can begin and one can see if there is indeed an impact on the outcome when using different rating methods. The average rating point comparison can be seen in Figure 4.3

4.2. Requirement Engineering Scenario

	Ranking	Categ.	Stars	Thumbs	Hearts	Slider
A	7,30	8,48	8,17	8,35	6,61	7,48
B	7,00	8,91	8,96	9,09	9,17	8,70
C	9,13	9,48	10,00	9,48	10,00	9,83
D	4,83	5,78	5,48	4,96	5,09	4,57
E	4,70	5,74	6,35	6,48	5,83	4,22
F	7,52	9,70	9,74	9,87	9,48	8,48
G	4,04	5,04	5,57	6,78	4,52	4,70
H	3,52	5,65	5,48	5,91	5,48	4,65
I	4,22	5,74	5,22	6,78	5,65	5,30
J	2,74	2,96	3,78	4,26	3,13	2,65

Figure 4.3.: Comparison of the average points between the different scaling methods

The columns represent the average ratings for a requirement regarding the different rating scales. The top ranked requirement per scale is marked with color to identify the winner at first sight. Also, one can see some sort of clustering. This happens due the conversion to the point system when the requirements are almost equally important to the user. Nevertheless, it still gives feedback about how participants prioritized those requirements. Despite the fact of possible contortions due the chosen requirements, there are some slight impacts identifiable. The ranking rating scale has a clear winning requirement and way less clustering between the top rated requirements than the other methods. The so to say losing requirements are also easier to determine. This is because people have to clearly think about prioritizing requirements with the ranking method because requirements can not have the same priority. Determining winners and losers is way harder in the other methods due heavy clustering in the top and bottom places except for the overall least rated requirement (J), which seems to do bad in any rating scale. With the first finding, one could expect that the

4. Evaluation

outcome of the slider rating method should not contain clustering due the high granularity of the scale. Most of the participants still rate different requirements equally even though they can easily give different ratings with the scale. So the high granularity slider scale has less impact than expected. Overall one can say that *the chosen rating method in requirement engineering scenarios has a slight impact on the outcome of a requirement engineering process (at least in this scenario), mostly on how clear the prioritization is.* So we can say that Hypothesis 1 can be confirmed to some extend, although the same requirements are always ranked top, middle, and bottom. This means the top, middle, and bottom cluster (if heavy clustering appears) usually consists of the same requirements which can be seen in the rank comparison in

Figure 4.4

	Ranking	Categ.	Stars	Thumbs	Hearts	Slider
A	3	4	4	4	4	4
B	4	3	3	3	3	2
C	1	2	1	2	1	1
D	5	5	7	9	8	8
E	6	6	5	7	5	9
F	2	1	2	1	2	3
G	8	9	6	5	9	6
H	9	8	7	8	7	7
I	7	6	9	5	6	5
J	10	10	10	10	10	10

Figure 4.4.: Comparison of the resulting ranks between the different scaling methods

There are some other factors which also should be taken into account when choosing a rating scale in your requirement engineering scenario. The investigation of those factors is described in the following subsections.

4.2.2. Tendency for Graphical Rating Scales

Now the focus lies on the three graphical rating scales in the study. One wants to know if people give higher or lower ratings when they use different graphical scales. For this, we take the average of over all votes from every participant in one rating scale and compare the value with the other scales. It is possible to compare the thumbs up/down rating to the stars and hearts because it corresponds to a one to five scale.

The average ratings regarding stars, hearts, and thumbs are:

- Stars: 3,88 (SD: 1,22)
- Hearts: 3,68 (SD: 1,31)
- Thumbs: 4,03 (SD: 1,21)

This is a really interesting and important finding. People tend to give higher ratings when using a thumb rating scale than using a 5-Star rating scale and they tend to give the lowest rating with hearts. This might be because of emotional aspects of the heart rating scale. Hearts are more personal and that is the reason why people save on them. With this finding, we can say that Hypothesis 2 can be confirmed.

4.2.3. Time Comparison of Different Rating Scales

In terms of time needed to rate the scenario with a certain rating scale, one can see in Figure 4.5 that participants using the ranking rating scale need

4. Evaluation

the most time and participants using the thumbs scale need the least time for rating the requirements. Interesting to see here is that the slider scale, which also has a higher granularity, does not take more time to use than a standard 5-Star rating scale. Also the hearts and thumbs rating scales seem to need less time than the 5-Star scale even though they are all graphical scales from one to five. So the conclusion here is that people take a lot more time when ranking objects where they can not have equal ratings and actually need to give a specific ranking.

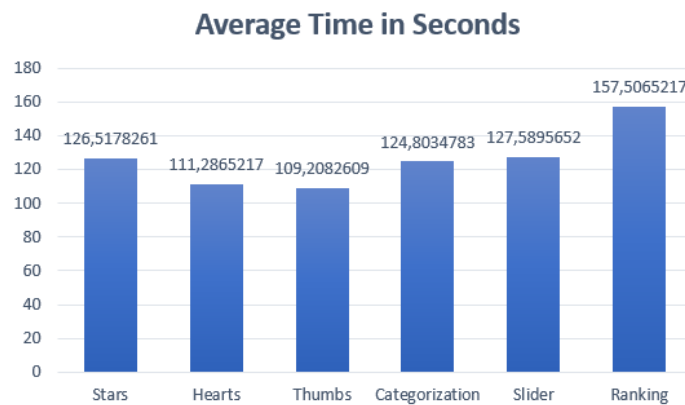


Figure 4.5.: The average time users need to rate all requirements with a certain scale

4.3. Usability and Understandability of the Rating Scales

To get an understanding on which rating scale has the best usability and which has the best understandability, we take the average value people give

4.3. Usability and Understandability of the Rating Scales

on usability and understandability for their used scale and compare them. In addition, one can evaluate the text-based comments to better understand why people give a certain value. The assumption is that there are rating scales that have better usability and understandability. Also that usability and understandability correlates.

The rating scale with the highest usability is a 5-Star rating scale. Participants who had to use the 5-Star scale in the scenario, give the highest points to that with an average of 4,39 (SD: 0,89). In Figure 4.6 one can see a graphical evaluation of the usability feedback.

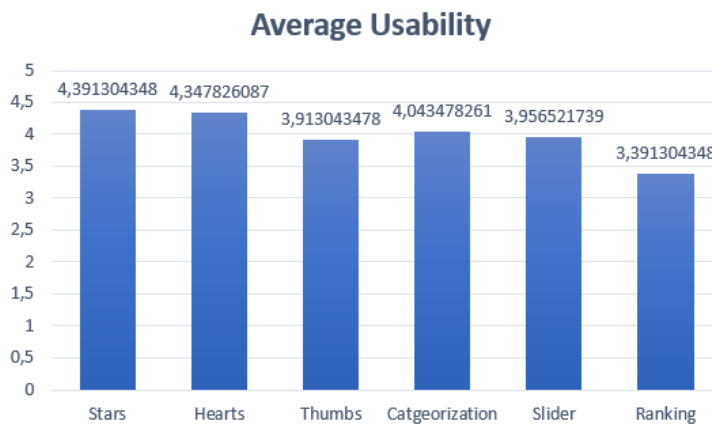


Figure 4.6.: Comparison of the average *usability* value between the different scales

The assumption that there are rating scales with a better usability was correct, although the difference between the average values is small. This result might not be significant due to the relatively small sample size and the small spread in the data. Due to the data, we can only distinguish between good and very good rating scales in terms of usability.

4. Evaluation

In terms of understandability, the winner is the hearts rating scale with an average value of 4,83 (SD: 0,49). The most confusing one is the thumbs rating scale which is shown in Figure 4.7. Similar to usability, the feedback from the study participants also confirms the assumption that there are rating scales with better understandability. Likewise, the difference between the average values is small. Regarding the correlation of usability and understandability, one can see that if the usability has a high value, the understandability also has a comparatively high value. The exception is the slider rating scale with a comparatively high understandability but low usability. This might be because of the possible high granularity of the slider scale. Participants easily understand that how sliders work but using them takes more effort in comparison to other scales because one has to do more than one click to rate.

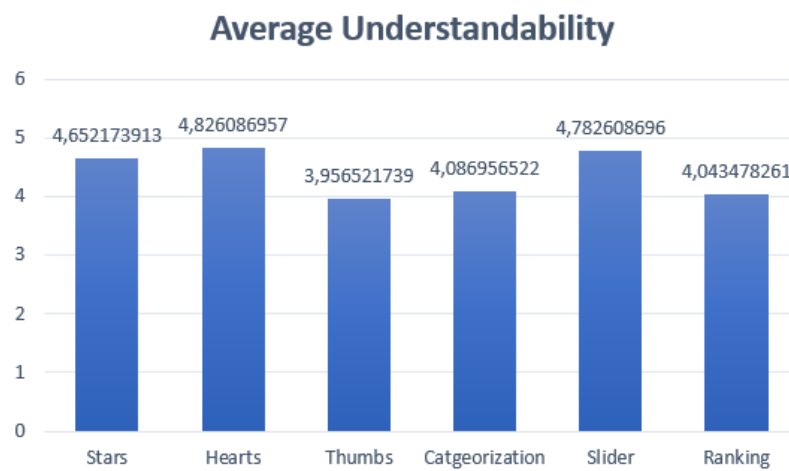


Figure 4.7.: Comparison of the average *understandability* rated by the users

4.3. Usability and Understandability of the Rating Scales

If we link those values to the time participants spent to rate with a certain scale, it is important to see that at least for the heart rating scale, which was the fastest scale, it holds that it is also the most understandable one and has a high usability value, but the thumbs scale, which is the second fastest scale, is the most confusing one. This might lead to the fact that participants just blindly click through the survey if they don't understand the scale and the survey allows this behavior. The ranking rating scale, for which the users take the most time to rate, also has the least points in usability and the second lowest points in understandability even though it gives the best outcome in terms of rating items in an order. So in general with some exceptions the time people spend relates to the complexity of the scale but the outcome also relates to the complexity of the scale which means that Hypothesis 3 was correct to a certain degree. This means there is trade-off between receiving useful data and let the user have an easy time to rate. These findings are probably domain-independent, because the complexity of a scale should not be linked with the question itself. Also, as described in Section 2.3, people take more time to rate with a more complex scale or even quit the survey earlier or regarding our case, quickly click through the survey if a scale is hard to understand, which indicates that the assumption that there are scales with higher understandability and usability and that those two factors correlate with the time effort is true to some extent.

4. Evaluation

4.4. The Most Preferred Rating Scales

4.4.1. Single Domain

To get information, whether there exists an overall preferred rating scale, we first evaluate the answers of the single domain questions. That means, we want to know if people like to use different rating scales for different attributes in one domain. An example would be "cars" as a domain and "design", "price", and "mileage" as attributes. The exact domains and attributes can be seen in Chapter 3. Looking at Figure 4.8, one can notice a tendency for the 5-Star rating scale.

Summary	5 Stars	Thumbs	Heart	Slider	Emotic.	Ranking	Categ.	Like/Dis.	Other
Domain 1-Attr. 1	11	3	1	7	3	10	9	1	1
D1-A2	9	3	2	10	6	5	9	1	1
D1-A3	12	4	3	7	6	7	5	1	1
D2-A1	13	2	9	3	8	2	4	2	0
D2-A2	12	1	4	10	8	2	3	4	2
D2-A3	13	5	5	4	6	2	3	6	0
D3-A1	16	6	4	5	11	4	4	6	0
D3-A2	15	6	11	3	8	5	6	6	0
D3-A3	10	7	3	9	14	5	5	3	1
D4-A1	19	4	1	12	2	4	6	2	0
D4-A2	10	5	3	6	13	1	5	3	0
D4-A3	8	2	2	7	1	7	8	3	0

Figure 4.8.: Outcome of the vote of the preferred rating scale in a single domain approach

Inside two out of the four domains we asked in the survey to rate for, there are some attributes which people would rather rate with a different scale than having a general rating scale for all attributes. Domain four represents the domain "Employee". This means we want to be flexible in terms of

4.4. The Most Preferred Rating Scales

rating scales when we have to rate other people. Other attributes where different rating scales would be preferred are the weather or the weight of a product. This clearly indicates that it is important to choose the right rating scale for your application also when there are different attributes inside a single domain.

The second question to answer is, if there is an overall preferred rating scale in the context of single domain questions. Analyzing the votes, the result over 276 possible votes for each scale in the single domain scenario looks like this (see Table 4.1):

5-Star	Thumbs	Hearts	Slider	Emoticon
148	48	48	83	86
Ranking	Categorization.	Like/Dislike	Other	
54	67	38	6	

Table 4.1.: Total votes for the rating scales in the single domain scenario

One can clearly see that in total, a regular 5-Star rating scale is by far the most preferred scale regarding the given domains and attributes. This is probably due to the high occurrence of 5-Star rating scales. This is the the first step of showing that there exists an overall preferred scale like we stated in Hypothesis 4.

4. Evaluation

4.4.2. Multi Domain

Now for the evaluation of the multi domain scenario of the study one wants to investigate if participants want to use different scales for different domains. An example would be "flowers", "boats", and "printers" as domains and a belonging attribute for each domain (e.g., the price for boats). The point is that the participants name their preferred scales for three completely different domains. The exact questions can also be seen in Chapter 3.

As shown in Figure 4.9, people like to have different scales for different domains. For reference it is important to know that each participant who got the multi domain question received three out of six possible questions. The three questions were the same for half of the participants. The other half received the other three questions.

Summary	5-Star	Thumbs	Hearts	Slider	Emotic.	Ranking	Categ.	Like/Dis.	Other
Q1-M1: "Songs"	12	8	9	3	5	6	4	12	0
Q1-M2: "Movies"	19	6	9	5	8	6	1	9	0
Q1-M3: "Games"	15	7	8	5	4	9	7	6	0
Q2-M4: "Logos"	14	3	4	6	8	11	6	4	1
Q2-M5: "Photos"	9	6	9	0	5	1	0	17	0
Q2-M6: "Papers"	12	1	2	11	4	8	7	1	1

Figure 4.9.: Outcome of the vote of the preferred rating scale in a multi domain approach

Although there are some domains (songs and photos) where people prefer a Like/Dislike approach, in general they choose the 5-Star rating scale as most preferred. This is probably because the 5-Star rating scale is the most

4.4. The Most Preferred Rating Scales

familiar one to the participants, which is detailed later in Section 5.1. Out of possible 138 votes per rating scale, the total number looks like this:

5-Star	Thumbs	Hearts	Slider	Emoticon
81	31	41	30	34

Ranking	Categorization.	Like/Dislike	Other
41	25	49	2

Table 4.2.: Total votes for the rating scales in the multi domain question

This is similar to the outcome of the single domain questions. So one could say that in total over both types of domain questions, the 5-Star rating scale is the most preferred and so the most convenient one to use for the users. As mentioned, there are some exceptions to this. Also we can be sure, that these findings are valid in terms of rating scales to choose from because the "other" option was hardly used. Overall we can say that Hypothesis 4 can be confirmed because the 5-Star scale is overall the most preferred scale.

5. Example User Interfaces

This chapter provides information about example user interfaces in real world scenarios and applications for the different rating scales we used in the first part of the study. This includes websites, apps and tools used for preference acquisition.

5.1. 5-Star Rating Scale

One of the most popular applications for 5-Star rating is the Amazon review page where people can rate their bought items which can be seen in Figure 5.1.

5. Example User Interfaces

Create Review



Huawei P9 Lite, Profer Tpu Protective Bumper Case Cover Ultra Thin Scratch-resistant Soft Flexible Silicone for ...

Overall rating



Rate features

Easy to hold ✕

Durability ✕

Value for money ✕



Add a photo or video

Shoppers find images and videos more helpful than text alone.



Add a headline

What's most important to know?

Figure 5.1.: A snippet of the Amazon review page where 5-Star rating is used

Other applications for a 5-Star rating scale are for example:

- Google reviews¹
- WhatsApp call reviews²
- Various survey creation tools (SurveyMonkey³, QuestionPro⁴, etc.)

¹www.google.com

²www.whatsapp.com

³www.surveymonkey.com

⁴www.questionpro.com

5.2. Thumbs Up/Down Rating Scale

Qiu, Parigi, and Abrahao (2018) mention the importance of 5-Star ratings and emphasize the universal usage, understandability and trustworthiness. Historical wise, 5-Star ratings were introduced for hotel classification. Taking a look on the star rating of a hotel became an important part in the booking process. (Denizci Guillet and Law, 2010) This familiar system has been brought to different domains and became the nowadays most used rating scale.

Despite all the positive aspects of 5-Star scales, there is a negative one about the ratings exemplified by Mukherjee, B. Liu, and Glance (2012), namely fake reviews. Common sense would tell that people would buy products with an average rating of five stars, but that is not the case in reality. People believe that products with many five star ratings have fake reviews because the product can't be *that* good. (Carbonell et al., 2019) This leads to a purchase likelihood peak when the average rating is between 4.2 and 4.5 according to a study by PowerReviews (2015). This means that shoppers are more likely to buy the item, if it's star rating ranges between 4.2 and 4.5, although this occurs more often for more expensive items than for cheap items which we also buy more often with a lower average rating.

5.2. Thumbs Up/Down Rating Scale

Objectively speaking, the thumbs up/down scale is one of the most important and recognized rating scale these days due to low complexity and the

5. Example User Interfaces

applications it is used in. It gives a more emotional result of the review than a standard scale because users "like" and "dislike" items. Using a thumbs up/down rating scale instead of a scale with higher granularity is better, if you want to provide the ability to quickly grab a user's opinion on an object, but is worse if you want to produce valid rankings. (**WhenThumbs**; C. Riedl et al., 2010) The thumbs up/down scale also be seen as a gamification method because they are a fun way to engage the community. Technical wise, one would use a thumbs up/down scale if a polarized result is more needed than degrees of opinion.

The biggest applications of a thumbs up/down rating scale are Youtube⁵ and Netflix⁶. Netflix even switched from a 5-Star scale to a thumbs up/down system. An article on the website "Business Insider⁷" outlines the reason behind the change. In the beginning of Netflix, they started out with a 5-Star rating. However, Netflix uses their rating system in a different way. Normally, if you go to a shopping website to buy a product, you will see the average rating of that product. Netflix now does not display the average rating over all users for a certain show, but tries to predict your rating. For that, Netflix uses a recommendation system. So when the user saw that a certain show has four out of five stars, they didn't see the average rating, but instead saw the prediction made by Netflix about how much you would like the show. So a four out of five meant that the user probably likes the show. This was very confusing to the users. To solve this problem, they

⁵www.youtube.com

⁶www.netflix.com

⁷www.businessinsider.de

5.2. Thumbs Up/Down Rating Scale

switched to a thumbs up/down rating scale and now display the prediction of how much a user would like a certain show to a percentage prompt which Netflix calls "Match". This can be seen in Figure 5.2



Figure 5.2.: Netflix show recommendation with rating scale and match percentage

Youtube also switched from a 5-Star rating to a thumbs up/down scale.(Sparling and Sen, 2011) This is not because people don't care enough to think about and granulate their rating. The problem is that most people don't do that. According to the data collected and provided on the official Youtube blog⁸, most users just gave one or five stars with nothing in between. This concludes that most users only rate the videos if they really like or dislike them.

⁸www.youtube.googleblog.com/

5. Example User Interfaces

This led to the decision to swap out the stars for a simple like or dislike thumb.

In Figure 5.3 one can see that the thumbs for rating a Youtube video are directly below the video itself. Also the user can immediately see how many other users liked or disliked the video.



Figure 5.3.: A snippet of a Youtube video page where Thumbs Up/Down is used

These days, the thumbs up/down scale is often used in multimedia systems, especially apps where you need a fast feedback whether the user likes something or not. Some example multimedia apps are:

5.2. Thumbs Up/Down Rating Scale

- Pandora Radio Station⁹
- Spotify¹⁰
- Dish Explorer App¹¹

A thumbs rating with a one to five granularity *like* described in the user study is used in the group decision app Choicla¹². This tool was indeed an inspiration for the study because of the good graphical aspect of the scale. In Figure 5.4, one can see how the thumbs are used inside the app.

The advantage of using thumbs instead of stars in the app is that the thumbs clearly indicate the state of the decision. The thumbs down options give the possibility for an overall dislike on an item where an overall “one out of five stars” rating just indicates that the item isn’t liked as much as the others.

⁹www.pandora.com

¹⁰www.spotify.com

¹¹www.dish.com

¹²<https://play.google.com/store/apps/details?id=com.selectionarts.choicla.android>

5. Example User Interfaces

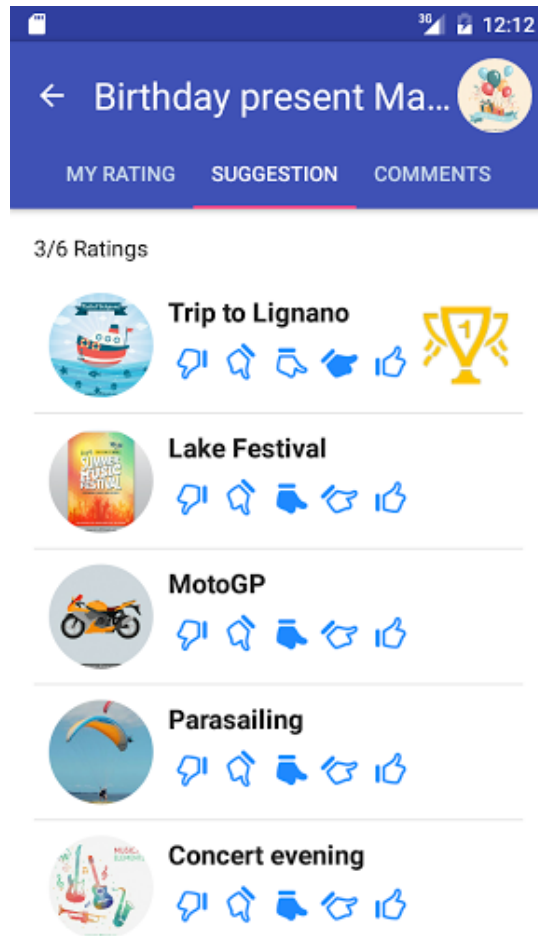


Figure 5.4.: A screenshot of the Choicla Android application

5.3. Ranking Rating Scale

Ranking rating scales are more of a niche technique to acquire preferences from users, even though they are used in several survey tools. The main

5.3. Ranking Rating Scale

applications are agile software development and requirement engineering. In requirement engineering, it is normally used as presented in Chapter 3. One big tool for agile development is Jira¹³ by Atlassian. The ranking rating scale is used to rank your issues (smaller task or requirements) via drag and drop. By ranking issues, users can actually arrange issues according to their relative importance (Scanlon, Christian, and Daily, 2019), this can be seen in Figure 5.5.

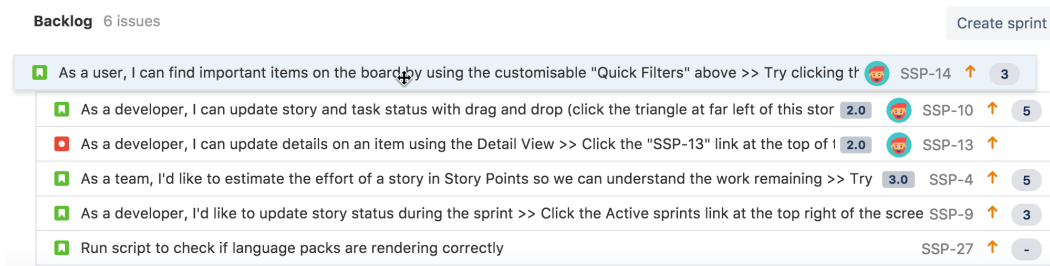


Figure 5.5.: A snippet of the ranking rating scale used in Jira presented on the Jira website

The same approach is also used in other agile tools like CA Agile Central¹⁴

The main difference between a standard rating scale (e.g., Likert Scale, Slider, etc.) and a ranking rating scales is that ranking scales have the possibility to force respondents to make a clear decision. (Harzing et al., 2009) Ranking enforces the user to identify which objects are most and least preferred. (Kalish and Nelson, 1991) Using a drag and drop system for the ranking also brings some benefits. Kunz (2015) mentions that by using a drag and

¹³<https://de.atlassian.com/software/jira>

¹⁴<https://www.ca.com/de/products/ca-agile-central.html>

5. Example User Interfaces

drop like system, one may prevent systematic response tendencies since respondents need to spend more time on it.

5.4. Heart Rating Scale

Unfortunately, there aren't many real applications which using this type of scale. This is kind of an interesting finding because like the scales before, rating scales using hearts are also available in various survey tools. These tools can often be customized, so that one can use their own images of stars and hearts instead of the provided ones. There are some artists and companies like Noun Project¹⁵ which provides different version of heart rating scales that can be implemented into your own survey. This shows that there is a need of having a heart rating scale available for your study. Most often, one wants to use hearts as a rating scale if they have an emotional affinity to the rated product. As one of the few applications of heart rating scales, we should mention personal blogs. People often create blogs to share their opinion on topics they like. Researching the internet for examples, we find that there are some personal blogs about books where people often use a self-designed heart rating scale to review and rate their last read books (for example Love Natalyn¹⁶) like shown in Figure 5.6.

¹⁵<https://thenounproject.com/>

¹⁶<http://www.lovenatalyn.com>

5.4. Heart Rating Scale

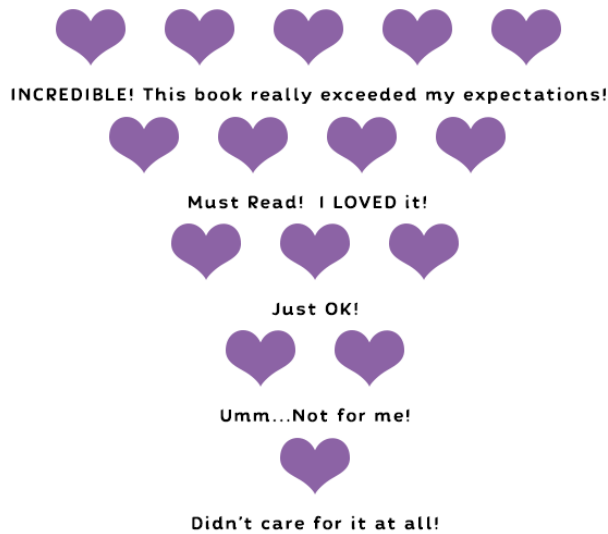


Figure 5.6.: The heart rating scale used on Love Natalyn¹⁶ for book reviews

Using hearts as a more emotional way to rate a product sounds good, but also brings some problems with it. Being emotional attached can bias your rating, which we found out in Section 4.2.2. On the other side, using hearts instead of stars also engages the user to rate more often. Another use case for hearts as a scale is Instagram¹⁷. On Instagram, one can “like” pictures. The “like” symbol for the button is a heart. If one picture receives more than 10 likes (hearts), the number of likes is also displayed. (Marr, 2018)

¹⁷www.instagram.com

5. Example User Interfaces

5.5. Slider Rating Scale

The slider rating is probably the most flexible scale in this study. Like the scales mentioned before, a slider type rating scale can also be found in many different survey tools such as Zoho¹⁸ and Qualtrics¹⁹. Depending on the used tool, slider ratings are using different start and end values. Most of the time, these values can be set manually. This also includes the step size if the slider is limited to a certain amount of steps. Another good example of such a slider is given by QuestionPro in their blog which can be seen in Figure 5.7.

Graphic Rating Scale

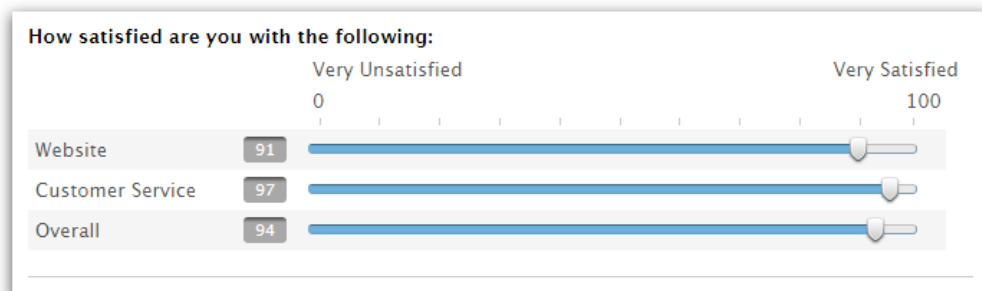


Figure 5.7.: Snippet of a slider scale in a fictive rating scenario

One would think that such a varying scale is used often in rating, reviews, and generally in preference acquisition systems, but that is not the case.

¹⁸<https://www.zoho.com/survey/>

¹⁹<https://www.qualtrics.com>

5.6. Categorization Rating Scale

Researching the web gives the conclusion that sliders are mostly used in configuration systems, for example RGB color slider. According to (Roster, Lucianetti, and Albaum, 2015), the argument for utilizing sliders is that they are less repetitive and more engaging for online survey respondents.

5.6. Categorization Rating Scale

This rating scale is mostly used for big questionnaires and surveys which are almost exclusively made with one of the different survey tools already mentioned before. One example for such a questionnaire is the "Game Engagement Questionnaire (GEQ)" where you have a many questions about a game you just experienced and a Likert-scale from one to five which correlates to "strongly disagree", "disagree", "Neither agree nor disagree", "agree" and "strongly agree". (Norman, 2013) In Figure 5.8, one can see how a categorization rating for an adapted GEQ looks like in Limesurvey (old design).

The categorization scale is also known as matrix scale. For survey creators, the matrix questions are easy to write and for respondents, the questions are easy to interpret since the answer options are the same across all items. (M. Liu and Cernat, 2018)

5. Example User Interfaces


Game Engagement Questionnaire: Please Answer all Questions

	1	2	3	4	5
I lose track of time	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I feel different	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I feel scared	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The environment feels real	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
If someone talks to me, I don't hear them	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I get wound up	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Time seems to kind of stand still or stop	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I feel spaced out	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I don't answer when someone talks to me	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I can't tell that I'm getting tired	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
My thoughts go fast	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I lose track of where I am	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The environment makes me feel calm	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I feel like I just can't stop watching	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figure 5.8.: Snippet of the Game Engagement Questionnaire

5.6. Categorization Rating Scale

Mobile clients often can't handle the horizontal formatting of the categorization matrix, so it is recommended to use a different layout, which can be seen in Figure 5.9



The image shows a mobile survey interface on a smartphone. The status bar at the top displays 'AT&T', signal strength, Wi-Fi, the time '3:38 PM', and a battery level of '50%'. Below the status bar, the text 'SurveyMonkey, Inc' is visible. The survey consists of three questions, each with five radio button options:

5. How satisfied or dissatisfied are your interaction with the sales staff?

- Very satisfied
- Somewhat satisfied
- Neither satisfied nor dissatisfied
- Somewhat dissatisfied
- Very dissatisfied

6. How satisfied or dissatisfied are the organization of the store?

- Very satisfied
- Somewhat satisfied
- Neither satisfied nor dissatisfied
- Somewhat dissatisfied
- Very dissatisfied

7. How satisfied or dissatisfied are you with the sizes available at the store?

- Very satisfied
- Somewhat satisfied
- Neither satisfied nor dissatisfied
- Somewhat dissatisfied
- Very dissatisfied

Figure 5.9.: A mobile layout for the categorization scale used by Survermonkey

5. Example User Interfaces

5.7. When to Use Which Scale

After looking into example user interfaces of our chosen rating scales, we should investigate on the most fitting scale for an application. In other words, find out "when should we use what". Starting off with the categorization rating scale (in particular the Likert-scale), this scale should be used if someone wants to gather quick feedback about multiple statements. If one wants to have a good prioritizing of products or features, a forced ranking scale is recommended according to Russell and Gray (1994). A graphical rating scale is in general a good idea to use in your system. According to Aggarwal and Mitra Thakur (2013) and Parill (1999), the advantages of using graphical ratings are user friendliness and a quick process. Star ratings are the most common used graphical scale, so if you don't want to overwhelm the respondents with an unfamiliar scale, this is the right *one* to choose. If you want to engage the users to rate more often, one wants to swap out the stars for a thumbs up/down system since this is a quicker way to rate something and the user doesn't get bored when thinking of a appropriate rating. If you plan to review products with an emotional affinity, one can swap out the stars for hearts. Finally, a slider scale should used if you want to give the user more freedom and allow them to select any value on a certain range. Buskirk, Saunders, and Michaud (2015) states that slider scales are good for mobile surveys because of the ability to slide on your phone display which is often not possible for computers unless they have a touchscreen monitor.

6. Future Work

After evaluating our user study in Chapter 4, we can see that there are some improvements that can be made for a future study and research on this topic. The requirement engineering scenario should be redone with different applications and the results should be compared with our findings. This will get us confirmation whether our evaluation results regarding the user behavior are useful. The chosen domain in this study is a very generic one and the given requirements could be too similar in terms of importance. Redoing the requirement engineering scenario with could also increase the overall significance of the outcome.

After those fictional scenarios, one should look into a real world requirement engineering process. This could be done in cooperation with companies using requirement engineering tools for developing products. The findings can be compared with the results of the fictional scenarios. This will also help to investigate the relevance of our results.

For the part of the study regarding the most preferred rating scale, a similar questionnaire could be done, but with more domains and attributes

6. Future Work

and with a better spread of topics. The domains can be put into different categories to get another research topic depending on a general best rating scale in a specific group of domains. Some of these categories could be: arts, technology, sports, and traveling. The findings here can be used, for example, on different shopping websites when evaluating bought items. Although one can research an algorithm for a better comparing method between different rating scales in an online shopping process. If person A wants to rate with stars (which could be automatically chosen by the system or configured manually), but person B wants to see ratings in a 1 to 10 scale, an algorithm could convert them to fit the persons preferences.

A big research question for the future is, how voice controlled devices like Amazon Echo¹ would influence us, if they are used to rate items. With such a device, the visual aspect while rating an item would completely be gone. Going even further with this approach, it would be interesting to know, if an artificial intelligence can be developed for a smart device that listens to a group of people and rank or filter their preferences. This could have a big impact in requirement engineering. Looking at another smart device, namely smart glasses like Google Glass², a question would be if it is possible to create a shopping list based on previous purchases. The idea is that the smart glasses would acquire your preferences by tracking your eye movement. The glasses would know if look often at specific items and add those items to your recommended shopping list.

¹<https://www.amazon.in/Amazon-Echo-Smart-speaker-Powered/dp/B0725W7Q38>

²<https://www.wearvision.de/googleglass/>

7. Summary and Conclusion

Rating scales do (slightly) have an impact on the behavior of people. Also, most people prefer the same rating scale (5-Star rating) when they rate items or products.

This is the main outcome of our user study where one part focuses on a requirement engineering process and the second on scales in general when rating different attributes in various domains (for example a digital camera bought in an online shop).

The main influencing factors of ratings scales in an requirement engineering process are granularity, usability, and understandability.

Usability and understandability are tied together. Scales that have a good understandability also have a good usability regarding to the participants in the study. Rating scales with a higher granularity tend to take more time if they are used to rank requirements in an requirement engineering process but there are some exceptions to this. This could be due to people not caring enough if they don't like to use the rating scale. Also, it seems that people

7. Summary and Conclusion

don't make use of a very large granularity in a rating scale. This contradicts a bit with the assumption that a higher granularity always results in better outcomes of the requirement engineering process.

Focusing now on the graphical rating scales (stars, hearts and thumbs) in detail, the conclusion is that depending on which scale people use, they tend to overrate and underrate with certain scales. People gave on an average higher ratings with thumbs and gave on an average the lowest ratings with hearts. So this could easily be due to a psychological effect and confirms the assumption that hearts are way more emotional than thumbs and stars and so people give them away harder and only give full points if they really like it compared to a full points rating with thumbs where people give higher ratings more frequently.

Usability and understandability are tied together when it comes to preferring a rating scale. This means people like to use scales that are easy to understand and to use, even if those scales do not deliver the best results. To get the information about rating scale preferences, we asked the users about their preferred rating scale for different domains in the second part of the study.

In general, participants really liked to use 5-Star ratings in any kind of domain. There are some exceptions where we humans do like other scales (for example like/dislike for pictures), even for different attributes in the same domain, but those are the minority. Also it is interesting to know that participants had an "other" option where they could type in any other scale

they would come up with, but nearly no feedback on this question was provided in our study. Which could mean that people are pleased with the common rating scales that were provided in the study and are used in various systems these days.

Appendix

Appendix A.

Processed User Study Data

The evaluation results after processing the data are presented in this appendix.

Appendix A. Processed User Study Data

A.1. Demographic Data

A.1.1. Gender

M	M	M	M	M	M		
M	F	M	F	F	M		
F	M	M	M	M	M		
M	F	F	F	M	M		
M	M	M	M	M	M	Males:	112
M	M	M	M	M	M		
M	M	M	M	F	M	Females:	26
M	M	M	M	M	M		
M	F	M	M	F	M		
M	F	M	M	M	M		
M	F	M	F	F	F		
F	M	F	M	M	F		
M	M	F	M	M	M		
M	M	M	M	M	M		
M	F	M	M	M	M		
M	M	M	F	M	M		
F	M	M	M	M	F		
M	M	M	M	M	M		
M	F	M	M	M	M		
M	M	M	F	M	M		
M	M	M	M	M	M		
F	M	M	M	M	M		
M	M	M	M	M	M		

Appendix A. Processed User Study Data

A.1. Demographic Data

A.1.2. Age

24 23 24 25 20 22
22 26 24 23 23 21
20 24 26 22 24 24
22 23 22 31 20 22
24 21 24 23 24 20
24 20 23 20 22 22
22 22 26 22 21 30
21 25 23 20 22 23
30 23 21 26 23 25
20 21 23 21 22 25
24 22 21 25 25 21
24 25 24 25 23 23
27 23 22 28 22 31
24 24 22 24 24 24
22 20 27 26 24 21
25 25 24 24 21 25
33 28 24 21 23 21
25 24 24 26 25 22
14 24 25 24 25 21
35 20 29 21 23 21
29 25 28 26 26 23
38 23 17 25 20 21
31 25 35 25 27 37

Average Age: 23,92028986

Standard Deviation: 3,474750622

Oldest Participant: 38

Youngest Participant: 14

A.2. Point System

A.2.1. Votes Converted to Points

The following tables representing each requirement (A-J) on the left and the corresponding points on the right.

A.2. Point System

Stars											Sum	Avg	Place														
A	5	10	10	10	10	10	7	4	6	10	5	10	3	10	6	10	10	8	10	7	10	7	10	7	188	8,17	4
B	10	6	10	10	10	10	8	10	10	4	10	10	10	7	6	10	10	8	10	10	7	10	10	7	206	8,96	3
C	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	230	10	1
D	1	10	2	10	4	6	8	3	4	10	2	3	3	10	7	4	5	10	3	10	1	5	5	126	5,48	7	
E	10	4	6	5	4	6	4	6	10	6	10	6	3	5	1	5	10	3	10	7	10	10	10	146	6,35	5	
F	10	10	10	10	10	4	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	224	9,74	2	
G	10	4	6	4	10	2	6	7	4	4	6	2	6	6	5	4	2	10	6	10	4	5	5	128	5,57	6	
H	4	4	4	4	2	1	2	10	4	6	10	6	3	5	10	10	10	3	10	4	5	5	126	5,48	7		
I	4	6	2	4	4	6	4	7	6	4	6	5	2	6	2	10	5	10	6	10	4	2	5	120	5,22	9	
J	4	4	4	4	1	4	4	2	1	6	2	1	3	6	1	4	1	10	6	10	7	1	1	87	3,78	10	

Appendix A. Processed User Study Data

	Hearts																				Sum	Avg	Place			
A	10	3	2	1	10	10	4	10	4	6	10	5	5	10	2	4	6	10	8	5	10	10	7	152	6,61	4
B	10	10	10	10	10	7	10	10	10	10	10	10	10	6	10	4	9	10	7	8	10	10	10	211	9,17	3
C	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	230	10	1
D	10	6	6	6	10	4	6	2	10	4	1	10	3	6	2	2	2	3	6	5	6	6	1	117	5,09	8
E	5	3	6	5	4	7	2	10	10	10	6	10	2	3	4	10	9	1	2	10	6	2	7	134	5,83	5
F	10	8	10	10	10	10	10	10	10	6	10	10	10	10	7	10	9	10	10	8	10	10	10	218	9,48	2
G	4	8	6	10	2	4	4	4	5	4	4	2	4	2	3	7	6	3	5	6	3	6	2	104	4,52	9
H	4	3	7	4	4	2	10	2	10	1	6	4	7	10	10	6	6	5	6	8	3	6	2	126	5,48	7
I	4	6	6	5	10	7	6	3	4	10	6	4	7	4	7	10	6	3	6	3	3	6	4	130	5,65	6
J	4	6	1	1	1	1	2	2	5	1	4	3	1	5	1	4	2	2	10	2	1	1	6	72	3,13	10

A.2. Point System

Thumbs	Sum	Avg	Place
A	10 10 10 4 8 5 10 5 5 10 8 10 9 10 10 8 10 4 7 10 10 10 9	192	8,35 4
B	7 10 10 8 10 10 10 3 8 10 6 10 10 8 10 10 8 10 10 10 10 10 9	209	9,09 3
C	3 10 10 10 10 10 5 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10	218	9,48 2
D	3 1 2 1 8 1 4 7 5 10 5 10 7 2 4 4 10 10 3 1 10 4 2	114	4,96 9
E	7 10 10 2 10 1 5 10 1 2 10 5 4 10 8 5 10 2 4 10 10 3 149	6,48 7	
F	9 10 10 10 10 10 10 10 10 9 10 10 10 10 10 10 10 10 10 10 10 9	227	9,87 1
G	3 10 10 4 8 5 4 5 10 10 5 10 5 10 4 5 2 4 7 10 10 10 5	156	6,78 5
H	7 10 10 4 5 10 5 10 10 2 10 5 10 4 3 3 4 1 4 3 4 2 136	5,91 8	
I	9 10 10 4 2 10 7 1 10 8 10 5 4 10 3 5 10 7 5 3 4 9 156	6,78 5	
J	7 10 1 4 1 10 2 5 2 3 5 10 5 2 4 3 2 1 7 4 1 4 5 98	4,26 10	

Appendix A. Processed User Study Data

	Slider										Sum	Avg	Place																
A	10	8	10	6	10	9	8	10	10	8	6	2	5	7	4	2	10	10	10	5	10	7	5	172	7,48	4			
B	10	8	10	10	10	8	6	10	10	7	4	10	10	10	10	10	10	10	10	10	8	10	4	9	6	200	8,7	2	
C	10	10	10	10	10	10	10	10	10	10	8	9	10	10	10	10	10	10	10	10	9	10	10	10	10	10	226	9,83	1
D	3	6	6	7	4	7	7	7	1	1	3	10	5	2	2	6	6	2	6	4	2	10	1	4	105	4,57	8		
E	5	5	5	2	2	3	4	3	3	2	7	4	1	7	10	5	7	1	3	10	3	3	2	97	4,22	9			
F	10	10	10	8	10	6	9	4	10	9	9	8	10	10	10	7	7	10	7	10	10	8	3	195	8,48	3			
G	2	4	3	6	5	2	2	10	10	4	5	4	3	3	3	3	5	3	6	1	10	4	10	108	4,7	6			
H	4	3	10	1	1	5	4	2	5	6	1	6	10	5	2	8	5	2	5	4	2	6	10	107	4,65	7			
I	6	3	1	3	6	5	5	10	2	5	2	8	10	5	5	2	2	5	2	10	10	5	10	122	5,3	5			
J	1	3	2	4	3	1	1	10	4	1	3	1	4	2	1	4	3	5	1	3	1	2	1	61	2,65	10			

A.2. Point System

Ranking											Sum	Avg	Place									
A	8	7	7	9	8	8	7	7	10	6	9	8	9	8	3	9	168	7,3	3			
B	9	8	4	7	7	6	9	9	9	9	7	6	8	7	4	8	4	5	6	161	7	4
C	10	10	9	10	10	10	10	5	8	10	10	10	10	10	2	10	9	10	9	210	9,13	1
D	3	5	8	4	6	7	1	2	8	2	5	2	3	5	8	3	10	4	7	111	4,83	5
E	5	6	2	6	4	4	4	5	2	4	1	7	2	6	6	6	2	6	6	108	4,7	6
F	7	9	10	8	9	9	8	8	3	7	8	9	7	9	10	7	1	9	5	173	7,52	2
G	2	3	1	5	3	5	6	6	6	5	6	5	4	1	1	5	7	5	4	93	4,04	8
H	4	4	5	3	1	3	2	4	1	1	4	3	6	4	5	2	8	2	3	81	3,52	9
I	6	1	6	2	5	2	3	3	7	10	3	4	5	2	7	1	6	3	2	97	4,22	7
J	1	2	3	1	2	1	5	1	4	3	2	1	1	3	9	4	5	1	1	63	2,74	10

Appendix A. Processed User Study Data

Categorization																					Sum	Avg	Place								
	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10											
A	7	3	10	10	10	10	10	10	10	10	6	7	6	10	10	7	10	5	10	10	10	10	10	4	10	10	195	8,48	4		
B	10	3	10	10	10	10	10	10	10	10	10	10	10	10	10	7	6	10	10	10	6	10	10	6	10	10	3	10	205	8,91	3
C	10	3	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	5	10	10	10	10	10	10	10	218	9,48	2	
D	3	6	10	4	10	2	3	2	1	3	10	10	5	10	10	2	10	2	10	3	10	5	2	7	5	133	5,78	5			
E	7	6	10	10	5	10	4	5	6	7	5	4	1	7	6	1	3	6	3	10	10	3	3	132	5,74	6					
F	10	10	10	10	10	10	10	10	10	10	10	10	10	10	6	10	10	10	10	10	10	10	7	10	223	9,7	1				
G	3	10	4	5	2	5	3	5	4	7	5	4	5	3	6	10	2	6	6	5	4	7	5	116	5,04	9					
H	7	10	4	3	5	2	10	5	7	3	3	4	5	3	6	6	10	6	3	5	10	10	3	130	5,65	8					
I	7	6	4	2	5	5	10	6	4	3	1	5	10	7	6	5	5	3	6	5	10	7	10	132	5,74	6					
J	1	10	1	2	1	5	3	2	4	4	3	4	5	1	10	5	1	1	1	1	1	1	1	1	1	1	68	2,96	10		

A.2.2. Point Comparison

Ranking	Categ.	Stars	Thumbs	Hearts	Slider	
A	7,30	8,48	8,17	8,35	6,61	7,48
B	7,00	8,91	8,96	9,09	9,17	8,70
C	9,13	9,48	10,00	9,48	10,00	9,83
D	4,83	5,78	5,48	4,96	5,09	4,57
E	4,70	5,74	6,35	6,48	5,83	4,22
F	7,52	9,70	9,74	9,87	9,48	8,48
G	4,04	5,04	5,57	6,78	4,52	4,70
H	3,52	5,65	5,48	5,91	5,48	4,65
I	4,22	5,74	5,22	6,78	5,65	5,30
J	2,74	2,96	3,78	4,26	3,13	2,65

Appendix A. Processed User Study Data

A.2.3. Corresponding Rank Comparison

Ranking	Categ.	Stars	Thumbs	Hearts	Slider	
A	3	4	4	4	4	4
B	4	3	3	3	3	2
C	1	2	1	2	1	1
D	5	5	7	9	8	8
E	6	6	5	7	5	9
F	2	1	2	1	2	3
G	8	9	6	5	9	6
H	9	8	7	8	7	7
I	7	6	9	5	6	5
J	10	10	10	10	10	10

A.2. Point System

A.3. Graphical Rating Scales

A.3.1. Votes With Stars

A	B	C	D	E	F	G	H	I	J
4	5	5	2	5	5	5	3	3	3
5	4	5	5	3	5	3	3	4	3
5	5	5	2	4	5	4	3	2	3
5	5	5	5	4	5	3	3	3	3
5	5	5	3	4	5	5	3	3	1
5	5	5	3	2	5	1	1	3	2
5	4	5	4	3	2	3	1	2	2
4	5	5	2	3	5	4	1	4	1
3	5	5	3	4	5	3	5	4	2
4	4	5	5	5	5	2	2	2	4
5	5	5	3	4	5	4	4	4	3
4	5	5	3	5	5	2	5	4	1
5	5	5	3	4	5	4	4	2	3
3	5	5	5	3	5	4	3	4	4
5	4	5	4	3	5	3	3	2	1
4	4	5	3	2	5	3	5	5	3
5	5	5	3	3	5	2	5	3	1
5	5	5	5	5	5	5	5	5	5
4	4	5	2	2	5	3	2	3	3
5	5	5	5	5	5	5	5	5	5
4	5	5	2	4	5	3	3	3	4
5	5	5	4	5	5	4	4	3	1
4	4	5	3	5	5	3	3	3	1

A.3. Graphical Rating Scales

Appendix A. Processed User Study Data

A.3.2. Votes With Hearts

A	B	C	D	E	F	G	H	I	J
5	5	5	5	4	5	3	3	3	3
2	5	5	3	2	4	4	2	3	3
2	5	5	3	3	5	3	4	3	1
1	5	5	4	3	5	5	2	3	1
5	5	5	5	4	5	3	4	5	2
5	4	5	3	4	5	3	2	4	2
3	5	5	4	2	5	3	5	4	2
5	5	5	2	5	5	4	2	3	4
4	5	5	5	5	5	4	5	4	2
4	5	5	3	5	4	3	2	5	3
5	5	5	1	4	5	2	4	4	3
4	5	5	5	5	5	3	3	3	1
3	5	5	2	1	5	1	4	4	3
5	4	5	4	2	5	2	5	3	1
1	5	5	1	2	4	4	5	4	2
3	3	5	1	5	5	4	4	5	1
3	4	5	1	4	4	2	3	3	1
5	5	5	3	2	5	4	4	3	5
4	3	5	2	1	5	2	2	2	1
3	4	5	3	5	4	2	4	2	1
5	5	5	4	4	5	4	3	3	1
5	5	5	4	3	5	3	4	4	4
4	5	5	1	4	5	3	2	3	4

A.3. Graphical Rating Scales

Appendix A. Processed User Study Data

A.3.3. Votes With Thumbs

A	B	C	D	E	F	G	H	I	J
5	3	2	2	3	4	2	3	4	3
5	5	5	4	5	5	5	5	5	5
5	5	5	4	5	5	5	5	5	1
4	5	5	3	5	5	4	5	5	4
4	4	5	4	2	5	4	3	3	1
4	5	5	2	5	5	4	4	3	5
5	5	5	4	2	5	4	5	5	3
3	5	5	4	3	5	3	3	4	3
4	5	4	4	5	5	5	5	1	2
5	4	5	5	3	5	5	5	5	4
4	4	5	3	2	5	3	2	4	3
5	5	5	5	5	5	5	5	5	5
4	2	5	3	1	4	1	1	1	1
5	5	5	3	4	5	5	5	4	3
5	5	5	4	5	5	4	4	5	4
4	4	5	2	4	5	3	1	1	1
5	5	5	5	4	5	2	3	4	2
4	5	5	5	5	5	4	4	5	2
4	5	5	3	2	5	4	1	4	4
5	5	5	2	3	5	5	3	4	3
5	5	5	5	5	5	5	4	4	3
5	5	5	4	5	5	5	4	4	4
4	4	5	1	2	4	3	1	4	3

A.3. Graphical Rating Scales

Appendix A. Processed User Study Data

A.4. Timings

Stars	Hearts	Thumbs	Categorization	Slider	Ranking
71,13	73,33	56,5	74,89	118,11	77,86
99,51	108,27	57,91	45,81	83,42	145,15
93,04	63,59	62,08	121,94	123,81	109,9
167,14	104,54	74,93	114,03	154,18	189,14
65,29	91,77	120,47	73,05	63,77	106,45
74,62	86,06	144,61	101,42	182,22	127,59
115,77	135,53	108,01	135,41	111,51	177,48
98,07	79	105,69	138,81	55,49	266,72
69,26	63,85	114,94	218,73	97,88	271,37
77,9	77,55	79,63	86,4	115,67	150,03
108,3	106,97	120,56	152,79	289,55	100,74
91,32	120,51	99,62	89,12	166,59	508,49
84,99	244,78	258,41	121,22	138,04	338,71
135,72	169,81	60,12	132,78	104,73	117,92
162	133,08	79,31	68,52	108,39	107,61
90,72	131,53	140,92	164,11	78,42	80,92
89,85	122,24	86,35	195,52	88,1	115,37
139,57	125,92	156,71	111,23	204,03	57,87
90,84	148,26	95,82	72,83	123,86	129,07
31,19	87,54	91,17	86,5	102,7	112,36
796,27	58,86	75,48	81,77	183,29	64,2
73,36	116,62	87,36	168,7	84,22	80,81
84,05	109,98	235,19	314,9	156,58	186,89

A.4. Timings

Appendix A. Processed User Study Data

A.5. Usability

Stars	Hearts	Thumbs	Catgeorization	Slider	Ranking
5	5	1	4	1	5
5	4	5	5	2	4
4	5	3	3	5	3
5	4	4	4	5	3
5	5	4	5	5	5
5	4	5	1	3	2
5	5	3	5	4	2
5	3	5	4	4	2
3	5	1	5	5	4
2	4	4	5	4	5
4	4	3	5	5	3
3	5	5	4	5	4
4	5	4	5	4	4
4	3	2	2	4	2
4	3	4	5	4	3
5	5	5	4	5	3
5	4	5	3	5	3
5	4	4	4	4	2
5	4	5	4	2	4
3	5	5	4	3	4
5	4	3	3	2	4
5	5	5	4	5	4
5	5	5	5	5	3

A.5. Usability

Appendix A. Processed User Study Data

A.6. Understandability

Stars	Hearts	Thumbs	Catgeorization	Slider	Ranking
5	5	2	5	3	5
5	5	5	3	5	4
5	5	3	5	5	5
5	5	5	5	5	1
5	5	5	5	5	5
5	5	5	1	5	5
5	5	5	4	5	4
5	4	5	4	5	3
5	5	1	5	5	5
5	3	4	5	5	5
5	5	4	5	5	4
4	5	1	4	4	4
5	5	4	5	5	5
4	4	3	4	5	5
3	5	5	4	5	2
5	5	4	5	5	4
3	5	5	4	5	5
5	5	3	4	5	5
5	5	4	4	4	3
3	5	5	4	4	3
5	5	3	4	5	4
5	5	5	3	5	2
5	5	5	2	5	5

A.6. Understandability

Appendix A. Processed User Study Data

A.7. Single Domain

Summary	5-Star	Thumbs	Hearts	Slider	Emotic.	Ranking	Categ.	Like/Dis.	Other
D1-A1	11	3	1	7	3	10	9	1	1
D1-A2	9	3	2	10	6	5	9	1	1
D1-A3	12	4	3	7	6	7	5	1	1
D2-A1	13	2	9	3	8	2	4	2	0
D2-A2	12	1	4	10	8	2	3	4	2
D2-A3	13	5	5	4	6	2	3	6	0
D3-A1	16	6	4	5	11	4	4	6	0
D3-A2	15	6	11	3	8	5	6	6	0
D3-A3	10	7	3	9	14	5	5	3	1
D4-A1	19	4	1	12	2	4	6	2	0
D4-A2	10	5	3	6	13	1	5	3	0
D4-A3	8	2	2	7	1	7	8	3	0
Total:	148	48	48	83	86	54	67	38	6

A.7. Single Domain

A.7.1. Individual Single Domain Results

	5-Star	Thumbs	Hearts	Slider	Emotic.	Ranking	Categ.	Like/Dis.	Other
Total D1:	32	10	6	24	15	22	23	3	3
Total D2:	38	8	18	17	22	6	10	12	2
Total D3:	41	19	18	17	33	14	15	15	1
Total D4:	37	48	17	31	41	28	31	27	8

A.7.2. Definition of Single Domains

Domain 1: Camera

D1-Attribute 1: Effective Resolution

D1-Attribute 2: Weight

D1-Attribute 3: Price

Domain 2: Flat

D2-Attribute 1: Location

D2-Attribute 2: Price

D2-Attribute 3: Construction Quality

Domain 3: Destination

D3-Attribute 1: Accessability

D3-Attribute 2: Landscapes

D3-Attribute 3: Weather

Domain 4: Employee

D4-Attribute 1: Expertise

D4-Attribute 2: Working Attitude

D4-Attribute 3: Achievements

Appendix A. Processed User Study Data

A.8. Multi Domain

Summary	5-Star	Thumbs	Hearts	Slider	Emotic.	Ranking	Categ.	Like/Dis.	Other
Q1-M1	12	8	9	3	5	6	4	12	0
Q1-M2	19	6	9	5	8	6	1	9	0
Q1-M3	15	7	8	5	4	9	7	6	0
Q2-M4	14	3	4	6	8	11	6	4	1
Q2-M5	9	6	9	0	5	1	0	17	0
Q3-M6	12	1	2	11	4	8	7	1	1
Total	81	31	41	30	34	41	25	49	2

Appendix A. Processed User Study Data

Question Set 1

Multi Domain 1: Songs

Multi Domain 2: Movies

Multi Domain 3: Games

Question Set 2

Multi Domain 4: Logos

Multi Domain 5: Photos

Multi Domain 6: Papers

Bibliography

- Aggarwal, Ashima and Gour Mitra Thakur (2013). "Techniques of Performance Appraisal-A Review." In: *International Journal of Engineering and Advanced Technology* ISSN, pp. 2249–8958 (cit. on p. 66).
- Askalidis, Georgios, Su Jung Kim, and Edward C. Malthouse (2017). "Understanding and Overcoming Biases in Online Review Systems." In: *Decis. Support Syst.* 97.C, pp. 23–30. ISSN: 0167-9236. DOI: 10.1016/j.dss.2017.03.002. URL: <https://doi.org/10.1016/j.dss.2017.03.002> (cit. on p. 16).
- Bollen, Dirk.G.F.M. (2015). "Modelling user preferences in multi-media recommender systems." English. Proefschrift. PhD thesis. ISBN: 978-90-386-3983-3 (cit. on p. 10).
- Branting, Karl and Patrick S. Broos (1997). "Automated acquisition of user preferences." In: *Int. J. Hum.-Comput. Stud.* 46, pp. 55–77 (cit. on p. 8).

Bibliography

- Buskirk, Trent D., Ted Saunders, and Joey Michaud (2015). "Are sliders too slick for surveys? An experiment comparing slider and radio button scales for smartphone, tablet and computer based surveys." In: *Methods, data, analyses : a journal for quantitative methods and survey methodology (mda)* 9.2, pp. 229–260. ISSN: 2190-4936 (cit. on p. 66).
- Carbonell, Guillermo et al. (2019). "The impact of emotionality and trust cues on the perceived trustworthiness of online reviews." In: *Cogent Business & Management* 6.1. Ed. by Andreea Molnar, p. 1586062. DOI: 10.1080/23311975.2019.1586062. eprint: <https://www.tandfonline.com/doi/pdf/10.1080/23311975.2019.1586062>. URL: <https://www.tandfonline.com/doi/abs/10.1080/23311975.2019.1586062> (cit. on p. 53).
- Dastani, Mehdi et al. (2005). "Modelling user preferences and mediating agents in electronic commerce." In: *Knowledge-Based Systems* 18.7, pp. 335–352. ISSN: 0950-7051. DOI: <https://doi.org/10.1016/j.knosys.2005.05.001>. URL: <http://www.sciencedirect.com/science/article/pii/S095070510500047X> (cit. on p. 10).
- DeCastellarnau, Anna (2018). "A classification of response scale characteristics that affect data quality: a literature review." In: *Quality & Quantity* 52.4, pp. 1523–1559. ISSN: 1573-7845. DOI: 10.1007/s11135-017-0533-4. URL: <https://doi.org/10.1007/s11135-017-0533-4> (cit. on p. 10).
- Denizci Guillet, Basak and Rob Law (2010). "Analyzing hotel star ratings on third-party distribution websites." In: *International Journal of Contemporary Hospitality Management* 22, pp. 797–813. DOI: 10.1108/09596111011063098 (cit. on p. 53).

- Di Sorbo, Andrea et al. (2016). "What Would Users Change in My App? Summarizing App Reviews for Recommending Software Changes." In: *Proceedings of the 2016 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering*. FSE 2016. Seattle, WA, USA: ACM, pp. 499–510. ISBN: 978-1-4503-4218-6. DOI: 10.1145/2950290.2950299. URL: <http://doi.acm.org/10.1145/2950290.2950299> (cit. on p. 1).
- Drosos, Dimitrios, Nikos Tsotsolas, and P Manolitzas (2011). "The Relationship Between Customer Satisfaction and Market Share: The Case of Mobile Sector in Greece." In: *INTERNATIONAL JOURNAL OF ENGINEERING AND MANAGEMENT* 3, pp. 87–105 (cit. on p. 14).
- Freyne, Jill, Shlomo Berkovsky, and Gregory Smith (2013). "Rating Bias and Preference Acquisition." In: *ACM Trans. Interact. Intell. Syst.* 3.3, 19:1–19:21. ISSN: 2160-6455. DOI: 10.1145/2499673. URL: <http://doi.acm.org/10.1145/2499673> (cit. on p. 9).
- Fu, Bin et al. (2013). "Why People Hate Your App: Making Sense of User Feedback in a Mobile App Store." In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '13. Chicago, Illinois, USA: ACM, pp. 1276–1284. ISBN: 978-1-4503-2174-7. DOI: 10.1145/2487575.2488202. URL: <http://doi.acm.org/10.1145/2487575.2488202> (cit. on p. 1).
- Funke, Frederik, Ulf-Dietrich Reips, and Randall Thomas (2010). "Sliders for the Smart: Type of Rating Scale on the Web Interacts With Educational Level." In: *Social Science Computer Review* 29, pp. 221–231. DOI: 10.1177/0894439310376896 (cit. on p. 11).

Bibliography

- Gemmis, Marco de et al. (2009). "Preference learning in recommender systems." In: *In Preference Learning (PL-09) ECML/PKDD-09 Workshop*, pp. 41–55 (cit. on p. 8).
- Gena, Cristina et al. (2011). "The Impact of Rating Scales on User's Rating Behavior." In: *User Modeling, Adaption and Personalization*. Ed. by Joseph A. Konstan et al. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 123–134. ISBN: 978-3-642-22362-4 (cit. on p. 13).
- Guerrouj, L., S. Azad, and P. C. Rigby (2015). "The influence of App churn on App success and StackOverflow discussions." In: *2015 IEEE 22nd International Conference on Software Analysis, Evolution, and Reengineering (SANER)*, pp. 321–330. DOI: 10.1109/SANER.2015.7081842 (cit. on p. 1).
- Guzman, Emitza and Walid Maalej (2014). "How Do Users Like This Feature? A Fine Grained Sentiment Analysis of App Reviews." In: *2014 IEEE 22nd International Requirements Engineering Conference, RE 2014 - Proceedings*, pp. 153–162. DOI: 10.1109/RE.2014.6912257 (cit. on p. 1).
- Harzing, Anne-Wil et al. (2009). "Rating versus ranking: What is the best way to reduce response and language bias in cross-national research?" In: *International Business Review* 18, pp. 417–432. DOI: 10.1016/j.ibusrev.2009.03.001 (cit. on p. 59).
- Hasan, Mohammad Shabbir et al. (2010). "An Evaluation of Software Requirement Prioritization Techniques." In: *International Journal of Computer Science and Information Security* 8.9, pp. 83–94 (cit. on p. 15).
- Holzbach, Rl (1978). "Rater bias in performance ratings: Superior, self-, and peer ratings." In: *Journal of Applied Psychology* 63, pp. 579–588. DOI: 10.1037//0021-9010.63.5.579 (cit. on p. 17).

- Kalish, Shlomo and Paul Nelson (1991). "A comparison of ranking, rating and reservation price measurement in conjoint analysis." In: *Marketing Letters* 2.4, pp. 327–335. ISSN: 1573-059X. DOI: 10.1007/BF00664219. URL: <https://doi.org/10.1007/BF00664219> (cit. on p. 59).
- Keusch, Florian and Ting Yan (2015). "The Effects of the Direction of Rating Scales on Survey Responses in a Telephone Survey." In: *Public Opinion Quarterly* 79.1, pp. 145–165. ISSN: 0033-362X. DOI: 10.1093/poq/nfu062. eprint: <http://oup.prod.sis.lan/poq/article-pdf/79/1/145/6863888/nfu062.pdf>. URL: <https://dx.doi.org/10.1093/poq/nfu062> (cit. on p. 11).
- Kunz, Tanja (2015). "Rating scales in Web surveys. A test of new drag-and-drop rating procedures." PhD thesis. Darmstadt: Technische Universität. URL: <http://tuprints.ulb.tu-darmstadt.de/5151/> (cit. on p. 59).
- L. Hughes, Jennifer, Abigail Camden, and Tenzin Yangchen (2016). "Rethinking and Updating Demographic Questions: Guidance to Improve Descriptions of Research Samples." In: *Psi Chi Journal of Psychological Research* 21, pp. 138–151. DOI: 10.24839/2164-8204.JN21.3.138 (cit. on p. 20).
- Linden, Greg, Brent Smith, and Jeremy York (2003). "Amazon.Com Recommendations: Item-to-Item Collaborative Filtering." In: *IEEE Internet Computing* 7.1, pp. 76–80. ISSN: 1089-7801. DOI: 10.1109/MIC.2003.1167344. URL: <http://dx.doi.org/10.1109/MIC.2003.1167344> (cit. on p. 2).
- Liu, Mingnan and Alexandru Cernat (2018). "Item-by-item Versus Matrix Questions: A Web Survey Experiment." In: *Social Science Computer Review* 36.6, pp. 690–706. DOI: 10.1177/0894439316674459. eprint: <https://>

Bibliography

- doi.org/10.1177/0894439316674459. URL: <https://doi.org/10.1177/0894439316674459> (cit. on p. 63).
- Marr, Bernard (2018). "The Amazing Ways Instagram Uses Big Data And Artificial Intelligence." In: *Forbes*. URL: <https://www.forbes.com/sites/bernardmarr/2018/03/16/the-amazing-ways-instagram-uses-big-data-and-artificial-intelligence/#38c747505ca6> (cit. on p. 61).
- Mauldin, Joshua (2014). "A Better Way To Request App Ratings." In: *Smashing Magazine*. URL: <https://www.smashingmagazine.com/2014/06/a-better-way-to-request-app-ratings/> (cit. on p. 1).
- Mukherjee, Arjun, Bing Liu, and Natalie Glance (2012). "Spotting Fake Reviewer Groups in Consumer Reviews." In: WWW '12, pp. 191–200. DOI: 10.1145/2187836.2187863. URL: <http://doi.acm.org/10.1145/2187836.2187863> (cit. on p. 53).
- Norman, Kent (2013). "GEQ (Game Engagement/Experience Questionnaire): A Review of Two Papers." In: *Interacting with Computers* 25, pp. 278–283. DOI: 10.1093/iwc/iwt009 (cit. on p. 63).
- O. Kingstrom, Paul and Larry E. Mainstone (1985). "An Investigation of the Rater-Ratee Acquaintance and Rater Bias." In: *Academy of Management Journal* 28, pp. 641–653. DOI: 10.5465/256119 (cit. on p. 17).
- Parill, Scott (1999). "Revisiting Rating Format Research: Computer-Based Rating Formats and Components of Accuracy." MA thesis. Virginia: Polytechnic Institute and State University (cit. on p. 66).
- Pinna, Sandro et al. (2003). "XPSwiki: An Agile Tool Supporting the Planning Game." In: *Extreme Programming and Agile Processes in Software Engineering*. Ed. by Michele Marchesi and Giancarlo Succi. Berlin, Heidelberg:

- Springer Berlin Heidelberg, pp. 104–113. ISBN: 978-3-540-44870-9 (cit. on p. 15).
- PowerReviews (2015). *From Reviews to Revenue How Star Ratings and Review Content Influence Purchase*. Tech. rep. Northwestern University's Spiegel Digital and Database Research Center (cit. on p. 53).
- Qiu, Will, Palo Parigi, and Bruno Abrahao (2018). "More Stars or More Reviews?" In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. Montreal QC, Canada: ACM, 153:1–153:11. ISBN: 978-1-4503-5620-6. DOI: 10.1145/3173574.3173727. URL: <http://doi.acm.org/10.1145/3173574.3173727> (cit. on p. 53).
- Rashid, Al Mamunur, George Karypis, and John Riedl (2005). "Influence in ratings-based recommender systems: An algorithm-independent approach." English (US). In: 5th SIAM International Conference on Data Mining, SDM 2005 ; Conference date: 21-04-2005 Through 23-04-2005, pp. 556–560 (cit. on p. 2).
- Ricci, Francesco and Quang Nhat Nguyen (2007). "Acquiring and Revising Preferences in a Critique-Based Mobile Recommender System." In: *IEEE Intelligent Systems* 22.3, pp. 22–29. DOI: 10.1109/MIS.2007.43. URL: <https://doi.org/10.1109/MIS.2007.43> (cit. on p. 7).
- Ricci, Francesco and Quang Nhat Nguyen (2005). "Critique-based mobile recommender systems." In: *ÖGAI Journal* 244 (cit. on p. 7).
- Riedl, Christoph et al. (2010). "Rating Scales for Collective Intelligence in Innovation Communities: Why Quick and Easy Decision Making Does Not Get it Right." In: *Proceedings of the International Conference on*

Bibliography

- Information Systems, ICIS 2010*. Ed. by Rajiv Sabherwal and Mary Sumner. Association for Information Systems (cit. on p. 54).
- Roster, Catherine, Lorenzo Lucianetti, and Gerald Albaum (2015). "Exploring Slider vs. Categorical Response Formats in Web-Based Surveys." In: *Journal of Research Practice* 11.1, Article D1. ISSN: 1712-851X (cit. on p. 63).
- Russell, P. A. and C. D. Gray (1994). "Ranking or rating? Some data and their implications for the measurement of evaluative response." In: *British Journal of Psychology* 85.1, pp. 79–92. DOI: 10.1111/j.2044-8295.1994.tb02509.x. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2044-8295.1994.tb02509.x>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2044-8295.1994.tb02509.x> (cit. on p. 66).
- Scanlon, Robin, Rhys Christian, and Laura Daily (2019). *Long-term agile planning with JIRA Software and Portfolio for JIRA*. Tech. rep. Atlassian (cit. on p. 59).
- Soat, Molly (2015). "Feeding the Addiction." In: *Marketing News* (cit. on p. 1).
- Sparling, E. Isaac and Shilad Sen (2011). "Rating: How Difficult is It?" In: *Proceedings of the Fifth ACM Conference on Recommender Systems. RecSys '11*. Chicago, Illinois, USA: ACM, pp. 149–156. ISBN: 978-1-4503-0683-6. DOI: 10.1145/2043932.2043961. URL: <http://doi.acm.org/10.1145/2043932.2043961> (cit. on p. 55).
- Thomas, Manoj and Ellie J Kyung (2018). "Slider Scale or Text Box: How Response Format Shapes Responses." In: *Journal of Consumer Research* 45.6, pp. 1274–1293. ISSN: 0093-5301. DOI: 10.1093/jcr/ucy057. eprint:

Bibliography

- <http://oup.prod.sis.lan/jcr/article-pdf/45/6/1274/28086417/ucy057.pdf>. URL: <https://doi.org/10.1093/jcr/ucy057> (cit. on p. 12).
- Wetzel, Eunike and Samuel Greiff (2018). "The World Beyond Rating Scales: Why We Should Think More Carefully About the Response Format in Questionnaires." In: *European Journal of Psychological Assessment* 34, pp. 1–5. DOI: 10.1027/1015-5759/a000469 (cit. on p. 15).
- Xiao, Bo and Izak Benbasat (2007). "E-Commerce Product Recommendation Agents: Use, Characteristics, and Impact." In: *MIS Quarterly* 31, pp. 137–209. DOI: 10.2307/25148784 (cit. on p. 2).