Towards Reliable Online Feedback : The Impact of User Preference and Visual Cues in Rating Scales and User Ratings

A thesis submitted to the
College of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements
for the degree of Master of Science
in the Department of Computer Science
University of Saskatchewan
Saskatoon

By
Maliha Mahbub

Permission to Use

In presenting this thesis in partial fulfillment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Disclaimer

Reference in this thesis to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement, recommendation, or favoring by the University of Saskatchewan. The views and opinions of the author expressed herein do not state or reflect those of the University of Saskatchewan, and shall not be used for advertising or product endorsement purposes.

Requests for permission to copy or to make other uses of materials in this thesis in whole or part should be addressed to:

Head of the Department of Computer Science
176 Thorvaldson Building, 110 Science Place
University of Saskatchewan
Saskatoon, Saskatchewan S7N 5C9 Canada

OR

Dean
College of Graduate and Postdoctoral Studies
University of Saskatchewan
116 Thorvaldson Building, 110 Science Place
Saskatoon, Saskatchewan S7N 5C9 Canada

Abstract

With the rise of dependency on online shopping and service providers, consumer ratings and reviews help users decide between good and bad options. Reliable and useful ratings can ensure better consumer service, product sales, brand management. Any underlying bias or external factors affecting users emotional stability can corrupt the credibility of user feedback. Prior studies suggest that the visual representation and design elements provided with a rating scale can affect the user's responses, specially if the rating scales have visual labels that generate an emotional response in users. Since there are a number of rating scale designs used in online e-commerce sites and recommender systems, it is also important that users get a say in which rating scale they are comfortable in using. Online marketplace still does not provide a platform to consider user's own choice in this matter. This preferential choice of scales can make users more involved in the rating process and help get the best response from them. Earlier research have already proved that users have specific personalized preferences when it comes to using rating scales to give feedback online. Further emphasis on how this preference and visual cues together can elicit more reliable online feedback mechanism is required in this area. This thesis aims to investigate whether the preference of users in rating scales influences the reliability and authenticity of user's ratings. It also explores the user's reaction to certain visual cues in rating scales, and how user's preferences of rating scale are influenced by such visual elements. A within-subject study ($n = 187$) was conducted to collect user ratings of popular products with six different rating scale designs, using two types of visual icons (stars and emojis) and colour-metaphors (using a warm-cool and a traffic-light metaphors). Statistical analysis from the survey shows that users prefer the scale with most visually informative design (traffic-light metaphor colours with emoji icons). It also shows that users tend to give their true ratings on scales they prefer most, rather than the scale design they are most familiar with. The rating score analysis also demonstrates a positive shift and better consistency in the ratings given on more visually rich scales. Based on these results, it can be concluded that user involvement is desirable in selecting the rating scale designs, and meaningful visual cues can contribute in getting more accurate (truthful) rating scores from users. The proposed approach of user preference based rating system has novelty because I elicited the user's own opinion on what their accurate or "true" rating is; rather than only relying on analysing the data received from the rating scores. This work can offer insights for online rating scale designs to improve the rating decision quality of users and help online business platforms obtain more credible feedback from customers which can significantly improve their services and user satisfaction.

Acknowledgements

I am grateful to almighty Allah, the most beneficent, the most merciful, for His relentless kindness, blessings and favours. I would like to express my gratitude to the best Supervisor anyone can ask for, Prof. Julita Vassileva. Her support, guidance and wisdom led my thesis to completion. She has not only been a supervisor, but also a compassionate, friendly, cooperative mentor to me and I am incredibly lucky to have such an outstanding woman by my side.

I would also like to sincerely thank my thesis advisory committee: Prof. Li Chen (external), Prof. Zadia Codabux and Prof. Cody Phillips for their valuable feedback and support. I would also like to thank my friends and labmates in MADMUC lab, who have always been helpful to me. A special thanks to Kha Thai Le, Mehnuma Tabassum and Abdul-Hammid Olagunju who helped me when I was facing difficulties in research.

I would like to thank my parents, Mahbubur Rahman and Farhana Akhter for the sacrifices, support, prayers and selflessness they bestowed on me and my sister Sraboni Mahbub for being my rock. I am forever indebted to you.

A special appreciation goes to my husband Zonayed Ahmed, for his constant support and inputs into my academic endeavours, which made my journey easier. Lastly, anyone who has encouraged me to not give up and has taken care of me during this time, I am deeply thankful for your presence in my life.

Contents

List of Tables

List of Figures

1 INTRODUCTION

“Word of Mouth (WOM)”, is defined as communication between trading parties regarding the evaluation of products and services [9], has been an essential part of business transactions. The enormous rise of the internet in the past decade has made business vendors and scholars pay special attention to the growth of electronic WOM or eWOM in the online marketplace and e-commerce sites [79], [9]. Where traditional WOM is often also limited to closed social circles of customers, the online reviews for a product or service are collected from users all around the world [9]. In online marketplace, product ratings and reviews have more power than traditional marketing or advertisements. Since the internet allows users to directly express their experience about a purchase, this is more reliable for consumers and can be used to build trust between online vendors and buyers [19]. And this is not just limited to e-commerce for buying and selling products such as Amazon and eBay; the online rating and review is now a “must-check before use” option for airlines, hotel bookings, movies, restaurants, resorts- almost all aspects of modern urban life all over the world. For customers, checking out the online ratings and reviews before considering buying is not just an option, but an expectation form the online sellers [9]. Product ratings are now consumer-generated contents which are an integral part of e-commerce and recommender websites [66]. Product ratings reflect the quality of products according to user experience and are very effective in predicting product sales [66] and recommending similar products to users. It also helps build trust between online vendor and buyers [73]. There has been steady stream of literature dedicated to the economic effect of online feedback and how to engage users in giving reviews that truly reflect the quality and user experience regarding products. The essential medium for collecting user feedback online is the rating tools such as rating scales, comment boxes etc provided on the websites. There are different types of rating scales available for consumers to rate the products they have used. These rating scales can vary in terms of their visual presentation from site to site such as thumbs up/down for Netflix, 5-star for Amazon, circles for TripAdvisor and so on. Researches show that different designs of rating scale can affect the ratings given by users to a considerable degree. User’s preference for rating scale designs also vary depending on how informative or easy to use the scale is. Despite the importance of consumer product ratings, there is a shortage of literature on the design of the product rating environments [92]. Online sites and recommender systems do not have any option to let users chose their own rating scale designs. Both companies and consumers make decisions on the basis of user ratings, it is essential to understand the degree to which these ratings can differ from their real value [92]. Consumer generated product-ratings are taking over traditional advertising and marketing strategies. Therefore, online business community and user interface (UI) and user experience (UX) designers must also revisit their transition from general, one-size-fits-all approach of rating systems to a more customized, adaptive approach. Since the consumer reviews and feedback are the driving force of today's e-commerce and recommender systems, the rating scale and feedback mechanisms offered to consumers should also revolve around how they prefer to use them.

1.1 Rating Scale Designs and User's Preference

Rating scales are visual widgets that are used to collect quantitative input to a system from users [17]. Both ratings given on rating scales or textual reviews in form of comments from users are popular tools for collecting users experience and opinions on a product or service. While reviews are considered more detailed and insightful, they require more cognitive efforts from the reviewers and time to process for the vendors. However, in the case of online products or services, many users tend to ignore reviews and depend more on numeric ratings provided on rating scales [92]. The rating scales possess several design elements or features which makes them unique in terms of both visual and functional perspective. These differences can be the number of points on the scale (5, 7 or 100), the presence or absence of a neutral middle point of the scale (e.g. in thumbs up/down scale, there is no neutral point) or the colour scheme used in scale (monochromatic or different colours in two ends of scale). Each of these design elements can affect how the users give their ratings on the scale. This difference of rating scores on different rating scales can create an unwanted bias that does not reflect the true opinion of users [16]. For example, if a user is overall happy with a hotel service, but wants some minor improvement, she/he can put a 7 on 10 on a 10-point scale. But if the hotel website offers a binary rating scale with only thumbs up and down options, the user is forced to give an extreme opinion which helps neither parties involved. Therefore, each rating scale has separate design factors which influence the ratings given on them due to the design factors such as granularity, visual metaphor, midpoint etc. [17], [18], [29]. Yet online selling platforms such as Amazon, eBay or recommender systems such as Netflix, YouTube provide only one fixed rating scale to their users.

1.1.1 Visual Heuristics for Rating Scales

The design factors of rating scales number of points, colour, visual metaphor etc. influence rating scores, but each of these factors have different effects on users. User's emotional intensity while using rating scales can cause fluctuations in the rating scores they provide [95]. Participants of surveys have certain difficulties when using survey tools [91]. When rating scales use certain visual signals or "cues" to ignite certain response in users, the users tend to be more engaged in the rating process. For example, using labels such as "very poor" and "excellent!" at low and high end of a scale respectively can make users more aware of the distance between high and low scores and they can feel contrasting emotions [55] due to this signaling. The process of rating products also increases the cognitive load of users as a result of which, the users tend to use certain cognitive shortcuts or "heuristics" to rate the items [93]. In this case, the rating scales which offer some sort of visual cues to the users can act as a heuristic measure that the users use to give ratings. In other words, rating scales that provide meaningful signal or cues by using certain visual elements or labels to ease the cognitive load of users, can contribute to getting more meaningful ratings from users [91]. For instance, when users use a rating scale that has different colours at the ends of the scale, the rating tends to shift towards more positive ends [90]. Whether rating scales are fully labeled or only labeled at the ends also creates a visible difference in the rating behaviour of users [65]. Using different granularity, i.e. different size of rating scale is also a major factor influencing rating behaviour as well [4], [17], [23].

The use of effective visual designs for getting better ratings from users has potential to change user's rating behaviour. There is a wide gap in the research of user preferences in rating scale features and designs and how it can affect the ratings of users. As a step towards reducing this gap, this research focuses on a novel framework for user preference based rating scale model. The goal of this rating scale model is to encourage users to give their true (in the sense of unbiased) ratings for a product by using rating scales which they prefer. The rating scales should also be designed to use strong visual cues to help users give ratings which are true and consistent to how they feel about a product. The research in this thesis aims to make contributions in the area of e-commerce by developing a better rating scale solution for both buyers and sellers, by getting the best possible feedback from the consumers.

1.2 Problem Statement

Engaging users in online rating and review and building trust has always been an issue for online merchants and recommender systems [11]. Positive or negative online ratings play a huge role in [19] product sales as well as the trust between buyer and seller [43], [41]. Prior studies suggest that respondents often face difficulties in using rating scales and they assign meaning to the visual elements present in the rating scales [91]. Therefore, providing a customized and personalized rating system can enhance user's engagement in rating process. Rating scale visualization should help users in the rating process by reducing their cognitive load [61]. By providing meaningful visual cues such as certain icons on scale points(thumbs or emoji) and different colours to identify meaning of each point on a scale can help users visualize the weight of the scores that they are giving as well. Since users can interpret such visual cues and assign meanings to each cue, it is very important to understand what meanings they are assigning to these cues. However, none of the prior studies on visual heuristics of rating scales explore which visual cues stimulate users towards more honest or truthful rating scores. By "true rating" in this text I denote the hypothetical internal, truthful, honest user opinion of an item (product, service, etc.). Allowing users to express how they feel about a product is essential for both online vendors and consumers. Rating scale designs are often designed with default preset formats which can be too monotonous for users and often fail to even persuade users to give a rating. If users are allowed to chose their own rating scale design, it can help them be more involved in the rating process and put more effort into rating.

Most of the earlier studies have investigated the role of visual cues and user preference of rating scales on the users but they did not focus on whether the ratings from the scales were truly reflecting how users felt about the products they were rating. For instance, Cena et al. [17] used a rating scale model to study the preferential choices of users in terms of rating scales and how that translates into different ratings on same product, but does not provide any insight on how far these different ratings on different scales are from the true ratings of users or which scales actually depicted the true ratings of users. The work by Gena et al. [35] also asserts the importance of allowing users to chose their ratings scales to encourage participation in recommender systems, but does not specify how exactly the preferential choices of rating scale can improve system performance. This thesis aims to address both these issues: visual cues and user preference of rating scales and attempts to understand the interrelation among user's rating scale preference, visual cues present in the scale and the true rating scores. The end goal of the thesis is to investigate the viability of a user preference based approach instead of a "one-size-fits-all" approach for rating scales in current online platforms. To summarize, the following issues are being addressed in the thesis:

- The effect of certain design elements which have cognitive connotations such as colours, icons in rating scores.
- The effect of visual elements/cues on the preference of users in rating scale design.
- Effectiveness of giving users option to choose their preferred scale in rating process (in terms of getting true feedback) over the current rating system where users get fixed rating scale for giving feedback.
- The role of user preference in rating scale design for eliciting true ratings from them.
- The possibility of users giving consistent and less negative ratings based on presence or absence of visual cues in rating scales.

1.3 Thesis Overview and Outline

In order to theorize my proposed framework, I present three hypotheses and aim to answer 6 research questions. In order to collect the user data, I designed a web-based rating application where participants are asked to rate products of their choice with 6 different rating scale designs. I used two different visual icons: emoji and stars with contrasting effects and two different colour metaphors, *traffic-light* and *warm-cold* with a monochromatic neutral shade to identify the different effect of each visual cue on user rating scores.
The thesis outline is as follows:

⇒ In chapter 2, I discuss the background theories and literature which contributed to the motivation of in this research.

⇒ Chapter 3 provides a review of related research works that are relevant to this research. And lastly, the hypothesis proposal and research questions are stated.

⇒ Chapter 4 presents a detailed description of the research methodology, the approach and the process involved, followed by the workflow.

⇒ Chapter 5 discusses the process of data collection, data processing and data analysis for implementing and validating first hypothesis which investigates the effect of visual cues in user preferred rating scales.

⇒ Chapter 6 demonstrates the frequent pattern mining and empirical tool for investigating the second hypothesis which focuses on the true ratings of users on rating scales.

⇒ Chapter 7 analyses the statistical and empirical description of third hypothesis which focuses on the nature of rating score on rating scales having ample visual cues versus non visually informative scales.

⇒ In chapter 8, the discussion of results from each experiment and the contribution of the research in terms of research statement and hypotheses are stated.

⇒ In chapter 9, the discussion of the future work, limitation and summary of the implications from my thesis are stated.

1.4 Publications from the Thesis

The research presented in this thesis has been partially published in the following conference papers:

- **Maliha Mahbub**, Najia Manjur, Julita Vassileva. Towards Better Rating Scale Design : An Experimental Analysis on the Influence of User Preference and Visual Cues on User Response. In Proceedings of the 16th International Conference on Persuasive Technologies, Persuasive 2021, April 12-14, 2021. Springer LNCS.

- Najia Manjur, **Maliha Mahbub** and Julita Vassileva. Exploring the Impact of Color on User Ratings: A Personality and Culture-Based Approach. In Proceedings of the 16th International Conference on Persuasive Technologies, Persuasive 2021, April 12-14, 2021. Springer LNCS.

2 LITERATURE REVIEW

With the ubiquitous presence of product ratings and reviews in the current two-way online marketplace, any unwanted biases in ratings can affect the quality of service as well as sales [8] with potential consumers. There are only a handful of rating scale designs such as Likert and Semantic differential formats which are used by the User Experience designers. The comfort and cognitive effort for using the rating scale can play a significant role when it comes to rating objects online. However, in most e-commerce sites, the user experience designers (UX designer) do not take the users' choice of rating scale and certain design aspects of rating scales into account. Many studies have been conducted on identifying and mitigating online rating and review biases caused by cognitive manipulation and user interface design elements. The majority of these studies focused around data driven post-consumption solutions. Capturing the accurate opinions of consumers is tricky as response biases such as anchoring bias, self-selection bias, extreme response tendency are very much an inseparable part on consumer rating and reviews. Rating scales design decisions that make the rating process and rating output reliable for both the online vendors and customers present a challenge in the fields of User Experience (UX) design, Human Computer Interaction (HCI) and E-commerce. This chapter summarizes the relevant literature and their influence on the proposed analysis and experiment on using rating scale visual elements (colour and icons) to detect and adjust biases raised from rating the same product using different scales.

2.1 Online Feedback Mechanism: Ratings and Reviews

With the rise of business transactions and marketplace online and era of e-commerce, WOM has also moved from small social circle of friends and families to being freely available to everyone with a device connected to internet [10]. Users post their review of a particular product, movie, hotel and services (e.g. Uber, Skip The Dishes) on the online forums to be seen by millions of online shoppers. Such reviews and ratings can collectively be defined as electronic word of mouth or eWOM. For consumers shopping online, the online review is not just an option that they can pass over, but it's an expected part of their shopping experience. The role of online recommendation with both pre and post-consumption is demonstrated in figure fig:feedback.

Online feedback and consumer reviews have drastically altered the way consumers shop, playing bigger role than traditional sources of consumer information such as advertising [58]. The 2015 report from global trust in advertising shows that almost two-thirds (66%) global respondents say that they trust online ratings and reviews [1]. According to a recent 2018 survey from Power Reviews, 97% of customers consult reviews for

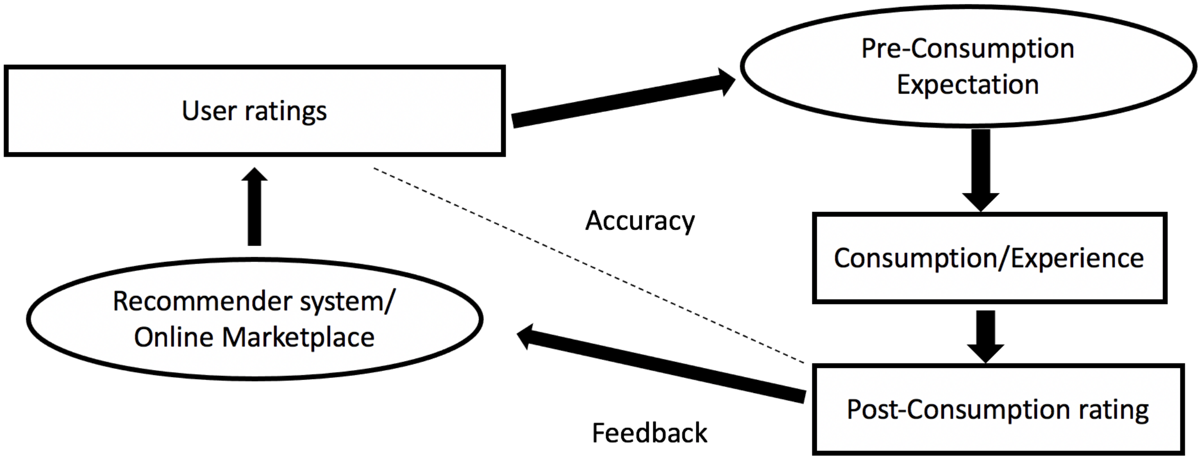


Figure 2.1: Product recommendation/sales and product rating, reproduced from [5]

every purchase they make, while 85% seek out negative reviews before considering purchasing any product. The same survey also found that 50% of consumers write reviews for products they've purchased, which is an increase from 42% in 2014 [2]. There is a large sector of literature that explored the economic importance of positive and negative ratings on sales as well as building trust between online customers and business platforms. The major areas studied were how online feedback mechanisms may help build trust between different parties in the online markets and how positive or negative ratings on sellers impact the final prices, the probability of a sale [78].

2.1.1 Role of Online Feedback Mechanism in Building Trust among Online Trading parties

Ba and Pavlou [11] proved that online feedback mechanism can induce trust among buyers and sellers which can mitigate information asymmetry among online transaction process. Since online customers have to rely on the information provided by the sellers without any physical inspection, there is a greater risk of falling for scams, fraudulent or misleading marketing gimmicks online than in traditional marketplace. Online vendors or sellers can easily maintain anonymity which makes such acts hard to track and bring to justice. Given the impersonal nature of online environment, building consumer-seller trust is much more crucial and complicated and this is where the requirement of online feedback mechanism plays a vital role. Online feedback mechanisms allow buyers to express their shopping/consumer experiences publicly by posting reviews and rating the quality of the service and products provided by the sellers. This feedbacks help build trust between the potential buyers and sellers in the online marketplace [96]. The experiment conducted by Ba and Pavlou [11] tested the effect of both negative and positive ratings on a buyers' purchasing decisions and willingness to pay premium price based on the feedback. The results showed strong evidence that positive feedback from buyers induces sustainable trust for sellers and this trust in turn can mitigate information asymmetry and garner price premium for reputed sellers. Building trust in e-commerce retail field has been investigated in detail by Resnick and Zeckhauser [78] by comparing the process of building trust between traditional retailers and online retailers such as eBay. Any bad transaction online is reported within seconds and can reach to millions of users in minutes. They introduced the reputation system for online sellers, which collects, distributes and aggregated past behaviour of consumers. Online feedback mechanisms and reputation systems have the same goals: building trust, customer acquisition and retention, product quality assurance, brand development, consumer service and so on [31]. Chrysanthos Dellarocas [31] published a detailed report on digitization of WOM and the differences between online feedback mechanisms and traditional word-of-mouth. The literature survey provided shorthands from other published works which evaluated the effect of positive and negative feedback on product sale and sellers' reputation as well as the trust between trading parties. The comprehensive studies of Resnick et al. [78] present and support the fact that positive feedback allows goods to be exchanged at higher prices. So customer ratings directly lead to economic benefit.

All these studies follow a similar trend which shows that the online feedback from users can have a substantial effect on the product sale, sellers' acceptability and overall online business platforms reputation. The work by Dellarocas [31] summarizes the findings from Resnick et al. [78] by stating following implications:

- Feedback results can affect both prices and the probability of sale for online business.
- The impact of feedback tends to be relatively higher for more expensive products and high risk transactions.
- The components of feedback information that seem to be affecting consumers buying decision most, are the overall number of positive and negative ratings, specially the number of recently posted negative reviews.

2.1.2 Role of Online Feedback Mechanism in Online Sales

Online feedback mechanism is essential for building trust and reputation between online trading parties. This trust and reputation has only one goal to reach eventually: increasing sales of online shopping communities. The trust and reputation of online vendors play the most influential role for online product sales. Chen et al. [19] focuses on the feedback provided on the products itself rather than on the sellers' quality. The empirical investigation on data from *Amazon.com* shows that an increase in consumer reviews and feedback can improve the sale of products, specially less-popular items can benefit more from the number of reviews.

As a result of such huge impact of online feedback on online product sale and seller reputation, online feedback collecting tools such as rating scales and review space are considered as an essential part of any online service provider. The online businesses which do not have any online rating or review collection system are considered highly risky and non -trustworthy by any user around the world and the role of previous user experience is the most important stamp of approval and branding conduit for online sales. Hu et al. [43] conducted qualitative and quantitative analysis which showed that good online reviews provided by customers can act as an catalyst for driving products sales upwards as users can differentiate between authentic good and bad reviews. Duan et al. [32] investigated the persuasive effect of reviews on movie box office sales, where online review and historical sales data are used to forecast movie ticket sales. Similar work based on forecasting product sales based on online reviews is investigated in Chong et al. [21]. The study found that product sales can be predicted from online reviews, promotional strategies and opinions. The effect of online reviews on electronic and video game sales have been investigated in [27], which showed the percentage of negative reviews having a greater effect than that of positive reviews. Li et al. [57] investigates the influence of both numerical and textual reviews on product sales and the findings have similar notions as negative reviews seemed to have stronger effect on creating negative impact on product sale than positive impact of positive reviews. The study also proposed that numeric ratings can mediate the effect of negative textual reviews. The effect of online reviews and ratings on eWOM and product sales have been proved to be substantial and influential on several studies [41], [48], [81]. In the study conducted by Hu et al. [42], the effect of both ratings scale and online review as feedback from users are analysed. The study developed a model to examine the relationships between ratings, sentiments and sales. The majority of these studies on online ratings and review analysis have illustrated significant findings which show direct impact of ratings and reviews on products sales. Such findings amplify the importance of a proper online feedback mechanism in every online marketplace. The challenges in the design and structures of online rating and review systems are discussed in the next section.

2.1.3 Challenges of Online Feedback Mechanism

The task of building a strong trust and reputation system for online marketplaces and its challenges have been studied in [60], which focuses on two -way trust building mechanism, where sellers can also give feedback on buyers. It also shows that users are more likely to leave a review or feedback if only they experienced extreme positive or negative experiences. This can potentially cause a distorted view of the actual service provided by the product and exhibit just the exceptional cases. Some literature on such challenges are stated below.

- Major challenges and some important areas of research in online feedback mechanisms have been stated in Dellarocas [31]. The authors studied and analysed the feedback mechanism of eBay to understand the process and properties of the system and listed some possible options to improve the performance of the feedback mechanism. They also compared the performance of online feedback with more traditional advertising and marketing strategies. One of the major challenges of online feedback mechanism listed in this study is eliciting sufficient and honest feedback. Ensuring sufficient feedback is challenging since users see no additional benefit from giving online feedback. Also, the majority of them prefer to wait unless a product has enough feedback before posting their own, no one wants to be the first to give ratings. At the same time, ensuring honest feedback is also a challenge since in most cases, users are rewarded or coerced to provide their feedback for an online purchase which may lead to biased ratings. Another study by Dellarocas shows that the success of online feedback system depends on number of participation and unbiased/truthful feedback [30].

- Another study on eBay's feedback mechanism shows the effect of the seller's reputation in online marketplaces [73]. The impersonal and anonymous nature of online marketplace requires a strong trust between buyer seller. There are two separate dimensions of trust on sellers: (a) Benevolence: The buyer's belief that a seller has goodwill and ethical motives beyond short-term profit motivations and (b) Credibility: The buyer's belief that a seller is trustworthy and competent, capable of fulfilling the transaction requirements on their part. The findings from the study showed that positive feedback comments highly increased buyers' trust both in terms of benevolence and credibility of sellers which in turn increased the price premium of items. When buyers trust a seller to be more credible and benevolent, they tend to pay higher prices and when they are uncertain of a seller, they demand monetary compensation in exchange for trust. These findings highlight the role of truthful online feedback for building a reliable buyer-seller relationship and successful online marketplace.

- Buyers are often reluctant to provide any negative feedback on the sellers since they want to avoid any harsh retaliation from sellers. The study by Li [56] states various issues with existing online feedback mechanisms such as untruthful reviews, a low incentive for providing feedback, and bias toward providing only positive feedback (since buyers want to avoid sellers retaliation). Li [56] proposes a mechanism to provide incentive to buyers for providing feedback to increase feedback participation and also to provide automatic feedback option to buyers in order to encourage honest feedback rather than only positive feedback.

- The marginal impact of both the positive and negative feedback on building trust in online marketplace has been analysed in [101] . This study showed that the impact of negative feedback is much higher than positive feedback on the final auction price in an online marketplace such as eBay. In all of the studies mentioned above, the online product price, profit and reputation of the seller are directly dependant on the online feedback participation and feedback nature (positive or negative) from the buyers.

One of the major challenges of online feedback and review systems is eliciting unbiased or truthful opinions and reviews from users. In most cases, consumers are not interested in giving feedback since there is no significant incentive for providing feedback for them. On top of that, online review and feedback mechanism face multiple biases: self-selection bias [58], social influence bias [86], response bias [38] and so on. During the early years of online shopping, the online review system was concerned about how to present the product qualities to users and feedback was solely based on user's experience with seller and product. With the exponential growth of online marketplace and online consumer-circle, users now start reading online ratings and reviews before even looking at the product description, making past reviews equally, and sometimes even more important in the decision making process. This can cause social influence bias and self-selection bias in user ratings. Social influence and self -selection bias for an online buying decision based on previous reviews have been the two major concerns for the scholars [9].

Social Influence Bias: Social influence bias can occur for popular products which tend to garner a large volume of positive or negative reviews from previous users. Many studies suggest that the motivation of a user to post a review does not only depend on their opinion about the product, but also on the existing reviews for that product as well [68], which contributes to social influence bias. If a certain item has mostly positive reviews (i.e., popular among users) then a user with negative experience may hesitate to post a negative comment due to peer-pressure. Muchnik et al. [67] showed that a random positive vote on a comment created a positive herding effect in a news website and the final ratings increased by 25% on average. Salganik et al. [80] created an artificial music market, where participants downloaded previously unknown songs, with or without the knowledge about the other participant's choices regarding the songs. The study found that the social influence played more important role in a song's success than song quality. If early reviews of a product are subjected to social influence bias, it can eventually pave way for self-selection bias.

Self-Selection Bias: Self-selection bias, can be stated as bias that occurs if only a selected set of users submit a review, which does not represent the general opinion of the entire consumer population for the product [28]. This type of bias happens when the early review and feedback of consumers are directed in e certain way (which can be either positive or negative) that do not match the review of general population over time. Therefore, the early responses of customers to a product can cause a self-selection bias [58] in subsequent reviews causing an anchoring effect in online reviews. For example, the fans of an author may post only positive reviews about a new book from the author while that book may not be well received by the general population. Thus, the positive review posted enthusiastically by the fans of the author does not reflect the truthful review of the book, resulting in a self-selection bias in rating.

Response Bias: The traditional customer surveys such as focus groups, brings together few people in one location to provide feedback for a product, service, idea or marketing campaign. Focus group participants are usually selected by the company themselves, typically unanimously. The participants are selected based on their purchase history, age, gender or personality. In contrast to this, the online customer review is not controlled by the company or website and they are voluntarily posted by customers, this can lead to unreported response bias [38]. This study demonstrates that customer's intention to post an online hotel review depends on the level of customer satisfaction. Customers posting online reviews are more motivated by extreme negative reviews than customers with positive reviews, which can cause underlying response bias. The study also reveals that when customers are familiar with the online review posting process, this unreported bias is mitigated. By simplifying review posting process for customers, the bias across the rating and tendency to post extreme and negative reviews can be reduced.

All these biases rising from online ratings and reviews can in turn create a volatile situation for online business and the seller-buyer trust and communication can hamper. Therefore, it is one of the major challenges for the online business community to tackle the challenges that come with online rating and review systems to ensure a transparent and open field for both trading parties in the long run. Since online reviews and their sentiment analysis can be multifaceted and more prone to bias than online rating [42], the easier and more effective way to elicit true and unbiased opinion of users about product quality is to ensure that the online rating of products and services reflect the true opinions of past users. The average online rating is also much faster and more convenient way for a consumer to judge a product rather than reading hundreds of reviews and comments [42]. The literature on consumer decision making process suggests that when users have abundance of information and options to chose from, they try to reduce their cognitive efforts by using simplified heuristics to arrive at a decision [13], [74]. Since a numeric rating provided on a rating scale requires less effort to process [83] than a textual review, consumers often resort to prefer using numeric ratings on rating scales to decide about a product as well as giving ratings about a product post-consumption.

Consumers have been able to see manipulation through ratings but not reviews [41] and numeric ratings have an easier grasp on user's decision making process [83]. Therefore, developing an online feedback mechanism to elicit true unbiased opinion of its users and exploring its role in shaping consumer decision making process are the focus of this thesis. The next section discusses the effect of online rating scale and it's design on the decision of online buyers and eventual product sales.

2.2 Effect of Rating Scale Visual Representation on User Ratings

Online rating scales are graphical widgets which allow users to express their opinions by means of a numerical score [35]. Online feedback collection process is simpler and faster when users use rating scale to express their views online. But biases discussed in previous section are also a part of responses collected from users who use rating scales. An added element that can influence user opinion while using rating scale is the design of rating scale and its features. There are numerous designs of rating scale that can vary in terms of colour, number of points, use of labels, etc. The rating scores of customers provided on different rating scales can vary depending on the design of rating scales.

The work by Cosley et al. [25] demonstrated that users can be manipulated by showing rating predictions when users rate movies and rating scales design. The study suggests that recommender system designers should focus more on rating interface design in order to deliver accurate "uninfluenced" recommendations. Reidl et al. [79] analyzed two frequently used rating scales: multi-criteria and single-criteria in online innovation communities to compare the outcome of using two different scales. The experiment results show that using a multi-criteria rating scale has improved the decision quality of users and their satisfaction with the use of scale than single-criteria scale. The study also revealed that when users have a pleasant experience using a rating scale, their decision quality and rating scores also improve significantly. Several studies [64], [40], [94] placed importance on the choice of appropriate scales for collecting user ratings. Multiple features on rating scales can have different effect on how users use the scale to give ratings they deem appropriate. Cena and Vernano [18] studied the difference of user opinions while using different rating scales and found that no single rating scale can have all the features and visual aspects to satisfy the specific needs and preferences of all users. Due to the heterogeneity in user preferences, using the same rating scale for all users could lead to misleading results and strained interaction with the system. Some users might find their experience with the system unsatisfactory and may decide not to interact with it at [18]. They considered two aspects of rating scales; the *granularity* and *emotional connotation* to study the influence of rating scales on the kind of ratings given by users. Specific connotations in the design of rating scales were identified from their previous experiment: thumbs are '"friendly" and "young" , but also "impolite". Stars are classical, familiar; sliders are precise, but "detached" and boring [18]. Their work provided the initial heuristic to adapt a system with customized rating scales based on users' preferences. In the work of Gena et al. [35], the rating scale features were identified as i) granularity, ii) numbering, iii) visual metaphor, iv) presence of a neutral position. Granularity means the numbers of points on a scale, which can be coarse (e.g. only having one negative and one positive point such as thumbs up/down) or fine e.g. 10 points slider scale. The numbering refers to the numeric labels attached to the ends or all points of scale such s -1, 0,+1 or 1,2,3,4,5 etc. The visual metaphors refer to the emotional or cultural connotation that the scale points can evoke in the users. For example, thumbs up/down is a human expression that is imitated in the scale; sliders are technical scale that imitates the measuring scales used in real life, emojis express real human facial expression of sad, happy, and so on. Neutral point is used to define a middle ground on the scale where users with no definite opinion can express their neutral opinion. The extensive experiments in [35] demonstrated that users exhibit different rating behaviour using scales with different features, such as thumbs up/down exerts a tendency towards higher rating, whereas a 3-point slider scale shows a tendency toward low ratings. The important result of the experiments is that most users appreciated the possibility/flexibility to choose the rating scale while rating items.

The research by Gena et al. [35] investigated the role of rating scales on user's rating behavior, showing that the rating scales have their own "personality" and thus using the same mathematical mapping from different rating scale scores is not ideal to elicit the accurate function of recommending items to users. According to [35], the rating scale designs can be are characterized by the fol- lowing features: i) granularity, ii) numbering, iii) visual metaphor, iv) presence of a neutral position. The features of rating scales investigated in this study are discussed below.

Granularity of Scales:

Sparling and Sen [85] evaluated the performance of scales with fine vs coarse granularity and concluded that users have different levels of satisfaction for using different scales and their cognitive load for using each scale also depends on the granularity. The user satisfaction level for using each scale ranging from fine granularity (sliders) to coarse granularity (unary) is shown in figure fig:satisfaction. The figure fig:satisfaction shows that the users dislike scales in the lower end of granularity spectrum (unary and slider), while strong likeness for five star scales. Similar premise has been investigated in more detailed framework in the work by Maharani et al. [61], where the the survey showed that the 5-star is still the most preferred and acceptable

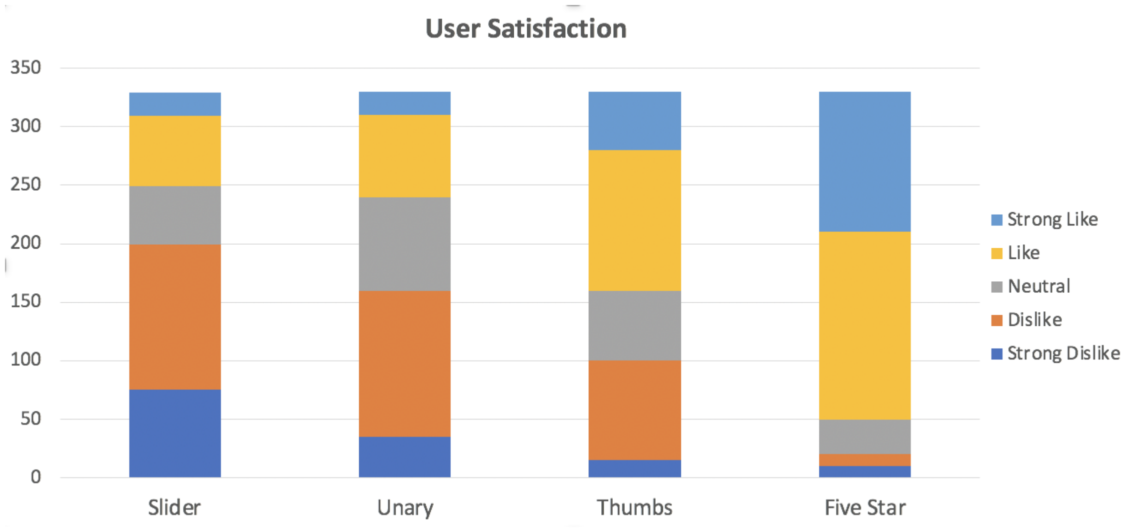


Figure 2.2: User satisfaction across all scales disregarding domain, reproduced from [85]

rating system for online ratings and reviews compared to thumbs up/down, unary rating, 5-star rating, a 10-point system and a 100-point system.

This work dives deeper into which factors played a role in choosing 5-star as the most popular rating scale visualization in terms of simplicity, readability and informativeness. In the experiment, most users (33%) chose 5-star rating as the preferred scale among other rating designs. Among the participants who preferred the 5-star rating, 32% stated that they chose 5-star rating cause it's simple, easy to read and understand, while 12% said that 5-star rating is familiar to them because it is widely used in commercial review websites. 8% of the respondents stated that 5-star rating is more informative and interesting and the rest of respondents stated that 5-star rating has clear criteria with eye-catching visuals [61]. The reasons for choosing each rating scale design is shown in figure fig:prefer. The following subsections discuss researches exploring different design features of rating scales and their effect of ratings given by users.

Neutral Point on Rating Scale:

The presence or absence of a neutral point in the rating scale can affect how users perceive the ratings they give on the scale. Lishner et al. [59] addressed the issue of limited scope for expressing opinions in traditional rating scales and proposed the implementation of rating scale that can adjust to the emotional valence of users. The experiment conducted by Adelson [4] compared the response of students to an additional neutral point in a 5-point Likert scale with a 4-point scale, which has no neutral midpoint. The participants preferred 5-point Likert scale with an additional neutral point along with clear negative and positive ends. Similar experiment conducted between 7-point and 5-point scale [23] illustrated that the response to 7-point scale is far more extreme compared to the 5-point scale, even though both has neutral point. Another experiment on consumer preference in [70] shows that exclusion of a neutral response option in the rating scale affects the judgment of extreme options. The study results showed that when users have mixed feelings about a

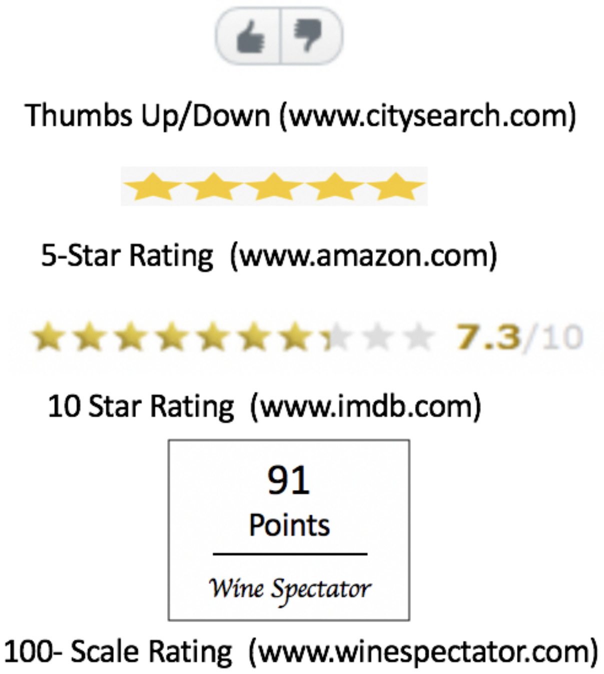

Thumbs Up/Down (www.citysearch.com)
5-Star Rating (www.amazon.com)
7.3/10
10 Star Rating (www.imdb.com)
91
Points
Wine Spectator
100- Scale Rating (www.winespectator.com)

Figure 2.3: Examples of rating granularity visualization, reproduced from [61]

product, their rating on 4-point scale with no neutral point is usually very different form the rating given on 5-point scale with neutral point. However, the presence of neutral point can also induce social desirability bias according to [34], as users may tend to select midpoint in order to keep everyone happy in survey process. The possibility of users choosing middle point as a mean to avoid being mean or harsh while rating has been extended in the hypothesis proposed in [92].

Numbering/Labelling of Rating Scales:

One of the most important factors that influences how consumers rate an item using rating scales, is the labelling on the points of scale. The rating scale points can be labelled with numbers such as 0, 1, 2, 3.... or the labels can be verbal such as “very poor” on low rating points and “Excellent!” on highest point on scale. Different types of labelling can have different impact on users rating choices while using the scales. The work in [71] stated that numerically labeled scales (0, 1, 2, 3, 4) may lead to more extreme responses than verbal scales (disagree to agree scales). The rating scales can be fully labelled , only end point labelled or not labelled at all as shown in figure fig:label .

Another study in [97] showed that whether scales are fully labelled i.e. all the rating points are labelled or partially labelled, i.e. just the endpoints are labelled can effect the response of users . When all points of a rating scale are aptly labelled with their apparent meaning or value, this can make the rating process easy for users, which reduces the extreme response tendency. On the other hand, when only the endpoints

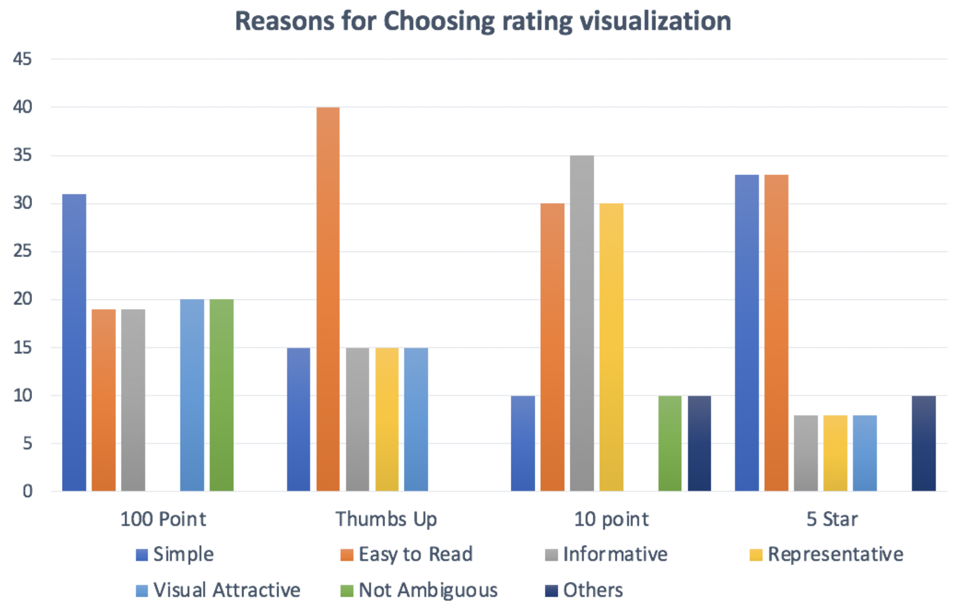


Figure 2.4: Reasons for choosing each rating scale, reproduced from [61]

of a rating scale are labelled, users struggle with rating process, resulting in extreme ratings for same items. Users also assign certain emotional value to endpoints of scales and ignore the intermediate labels [100]. This persuasive power of emotional cues caused by labelling the rating scales can affect how users rate products as well which can steer them away from their true ratings [98]. Also, since emotional labels ignite the active thinking process of users, they become more engaged in the rating process and extend their range of responses. Users tend to give extreme and random ratings when using scales which are not labelled, since these scales have no emotional cues attached to them (as the values given on scales feel arbitrary to them) [92]. On a different note, fully-labelling all points of rating scales can increase the cognitive effort of processing the rating options [52]. Labelling every points distinctively in the scale can enhance the reliability of the rating scores on such scales [7]. The majority of studies show that acquiescence is higher and extreme response bias is lower with fully-labelled scales compared to only end-point labelled or non-labelled scales [65]. Keeping with such findings, different effect of verbal, numeric and visual labels in rating scales is investigated in both [88] and [90]. The findings in both of these studies show that the effect of verbal cues (such as agree, disagree, poor or excellent) is higher than using numbers on scales. When participants were given fully labelled scale with different shaded end points, they tend to chose positive reviews.

Visual Icons used in Rating Scales:

The use of pictorial images or visual icons such as emojis or hearts instead of number or generic scale points (radio buttons or slider) on rating scale can influence the rating score given on the scales. The impact of using labels on rating scales have already been studied and the effect of different types of labeling on user ratings is analysed in subsequent studies. The focus of these studies was to see how visual representation of rating scale instead of generic rating points influence user rating scores. The experiment conducted by

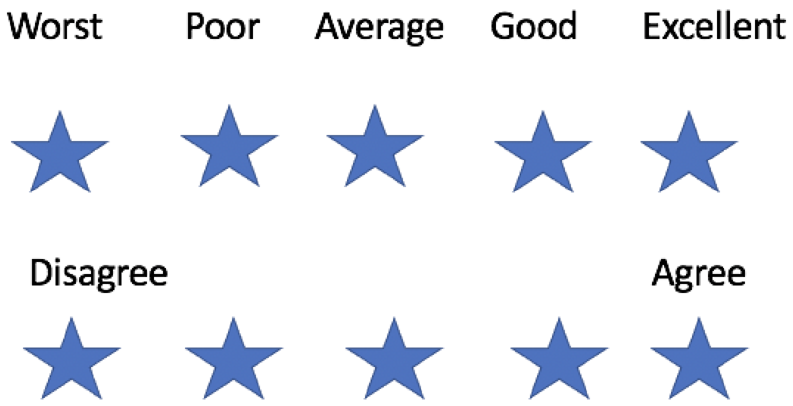


Figure 2.5: An example of fully labelled and endpoint labelled scale

Toepoel and Dilman [88] showed the presence of a hierarchy of words, numbers and visual cues/ labels in rating scales. The results from their experiment showed that visual labels such as emoji and colour on rating scales and verbal labels can cause a shift towards positive ratings. Adding numbers or numeric labels to scales can diminish such effects, causing users to rate more conservatively. A more detailed study was conducted by Vera Toepoel et al. [89] by replacing the numeric /measurement scales with pictorial icons on rating scales such as hearts, smileys, stars, button, grid or tiles as rating scale points. The results showed that hearts and stars had lower average scores from users and grid designs were the most negatively evaluated. Respondents evaluated the smiley design most positively. This study contributed to significant findings in terms of the use of visual icons instead of generic rating scale points. The findings showed that using visual icons as rating scale points can yield significantly different results than using general scale points.

2.2.1 Classification of Rating Scale Design

An important study evaluating effect of using different rating scale designs on user rating is conducted by Cena and Vernero [17] which analysed and classified different features of rating scales based on based on granularity and visual metaphor. Using icons such as thumbs, hearts or stars instead of generic numeric labels had a significant effect on how users view the rating process and the rating score they give on such scales. They devised a rating scale model based on intrinsic and extrinsic features of rating scales in previous studies which is shown in figure fig:cenamodel.

The user study attempts to figure out which scale would the users choose to rate different object, whether they prefer different scales for evaluating different objects, whether user's choices change after rating repeatedly (by gaining more experience on evaluating an object) and the motivation for user's choice of scale. The classification of rating scales by Cena and Vernon [17] identified three major classes of rating scales based on

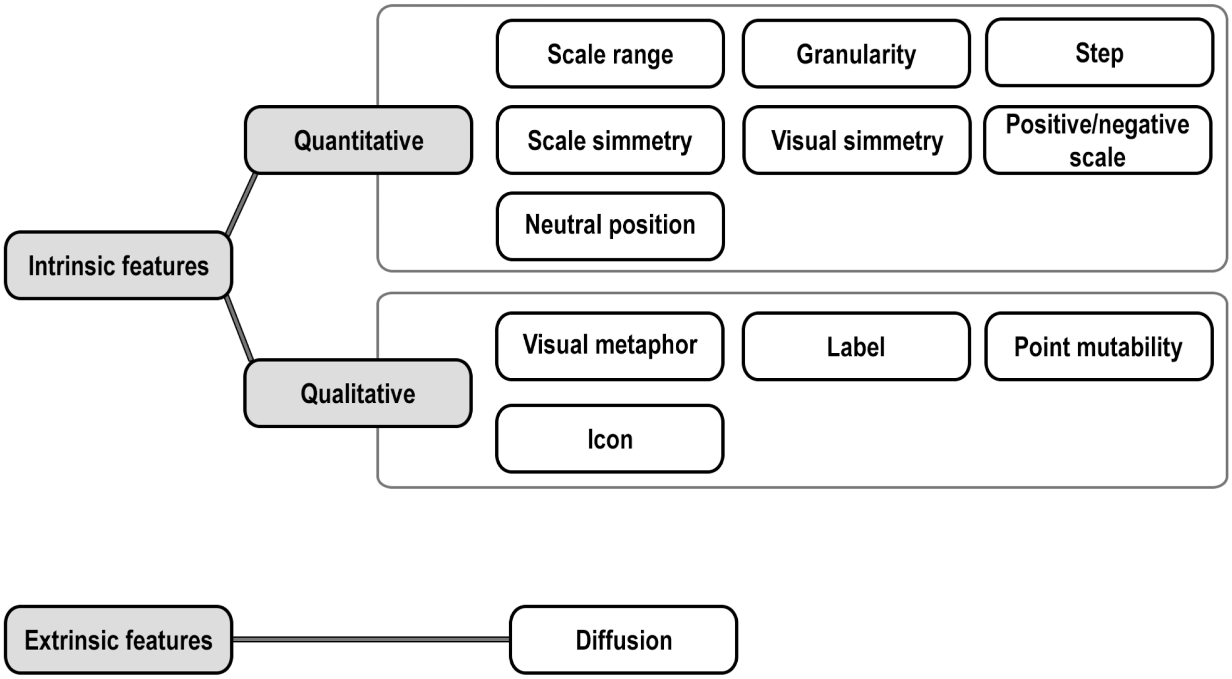


Figure 2.6: Rating scales features identified from previous works, reproduced with permission from [17]

12 intrinsic and extrinsic features of rating scales deduced from previous works by [94], [35] and [69]. The authors have conducted extensive user study to identify preferential choices about rating scales in a website. These features are sorted into two major groups and each feature is analysed and finally they were used to classify rating scales into three clusters. The features and their groups are presented in table.

The three major classes of rating scales identified in this work are:

- Human Scale: Scales which has a strong visual aspects which exploit human emotions to connote meaning for each rating point on scale. A good example is the human face emoji.
- Neutral Scale: Scales which rely on quantitative inputs rather than visual characterizations. All the icons used in this scale are same and have no strong emotional/ visual connotations. They are often considered standards such as star.
- Technical Scale: These are scales which have a strong focus on quantitative evaluations, similar to those of measurement tools. They use no icons/image for their measurement points, rather rely on abstraction with no differences in the visual representation of scale points such as sliders/likert scales. The scale points have numeric labels to identify values on each point.

The authors have used three scales from each class: emojis for human scale, stars for neutral scale and slider for technical scale. The different types of scale used in their experiment is showed in figure fig:cenamode. The users liked the star scale due to familiarity and disliked sliders the most. The most important factors in

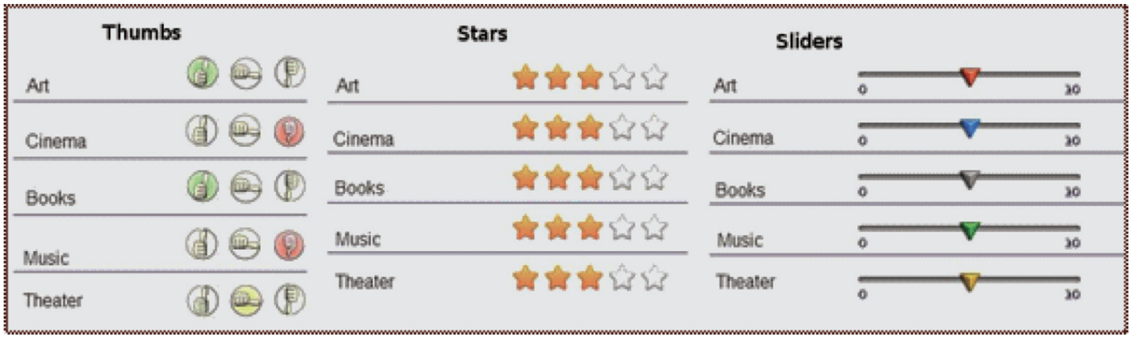


Figure 2.7: Rating scales features used in experiment, reproduced with permission from [17]. 3-point thumbs for the "human", 5-point stars for the "neutral" and 11-point sliders for the "technical" class

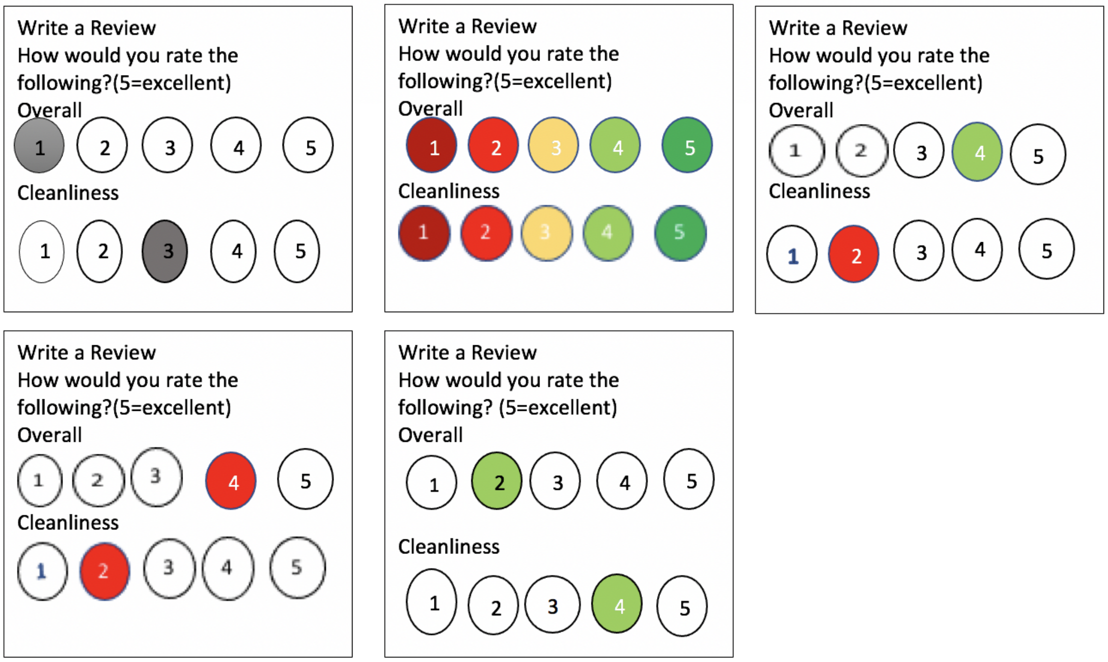


Figure 2.8: Experiment for different colour treatment in rating scales, reproduced from [14]

users' preference for rating scales were found to be granularity and visual metaphor. The authors suggested using separate rating scales in order to elicit the best rating performance from users. In addition to using different visual icons as rating scale design, use of different colour in rating scale has also been an area of study for user rating analysis. In the next section, the use of colour as a mode to influence the ratings of users is discussed.

2.2.2 Role of Colour Metaphor in Rating Behaviour of Users

Using colour as a cognitive tool to assert certain emotion in users has been a very widely regarded topic in Human Computer Interaction and e-commerce [15]. Certain colour metaphors are imprinted in our minds which can influence our emotions; such as seeing red as "danger" or green as "positive" or seeing black as “ominous” and so on. One of the most common colour schemes for triggering opposite emotional response in users are the red-green and red-blue colour schemes. These colour metaphors and their use in affecting user behaviour are studies in various literature ranging form marketing to human-computer interaction. One of such studies experimented the use of red and green to identify healthy and unhealthy food items in [54]. In this study, 20 popular food items were colour coded based on the calorie and general nutritional quality in them. The colour coded categories were most healthy (green), medium healthy (yellow), or least healthy (red). The red-yellow-green colour metaphor is also known as traffic light metaphor as this is the exact same scheme used in traffic light all over the world. This classification resulted in customers choosing much more healthier food compared to food items with higher calorie intake when they are coded with traffic light colour coding.

The use of red as exciting, warm colour and blue as relaxed and cool colour in marketing and logos has been explored in [53]. The same logo presented in different colour was related to different emotions on users, for example, the colour red, orange and yellow effected the excitement factor whereas the colour green, blue and brown in logos positively effected the sense of sincerity or calm in users. Different colour hues can relate to different emotional response in users which can affect how they react to the environment where the colour is present. The cognitive impact of color cues also depends on the user’s interpretation of what the colour signifies (stands for) in human mind due to constant exposure. This means when users are habituated in seeing a particular colour used for a particular context or purpose, the colour itself can stand for that context in their mind even when not used in same purpose. This has been called the signaling significance of the colour in human mind in [44]. For example, since traffic light uses red as a signal to “Stop” and green as a signal to “Go”, these two colours can have similar effects on human mind when used in different contexts. Which is why, red is often used for stop or off button in electronics and green is used as on/ start button.

The impact of colour on the users have been applied in rating scale designs in many literature. The traffic light metaphor is explored in the work by Bonaretti et al. [14] by using colour in rating scales. Here, the authors used red as a negative and green as a positive heuristic which is derived from the widely popular “traffic-light” metaphor in 5 different treatments and compared the rating provided with a baseline “grey” rating scale, as shown in figure fig:traffic to see how emotional intensity of colour can affect the rating bias in users. The authors hypothesized that using traffic light colour metaphor in rating scale can increase the emotional intensity of users which is eventually reflected in the ratings given on such scales. Similar colour-shading was used in experiment by Maharani et al. [61] who used red and blue as colour cues for negative and positive ends respectively. The idea for Maharani’s colour shading scheme is relatively simple, the colours were just used to see if different shading at two ends of rating scale help users visualize the extreme rating points in 5-star rating scales. The colour coded 5 -star rating was preferred by most users in the study due to its clear polarity in defining the extreme positive and negative ends.

A more detailed approach in this regard was adapted by Tourangeau et al. [90], where the authors used both different shades of same colour in the scale and different colours at different ends of scales to see the effect in users response. The results demonstrated that using different shades of colour at two ends of a scale can cause a significant shift in the ratings given by users on the scale. The colour scheme in this case was again red and blue(red being negative and blue at positive end). The study by Pilipczuk and Cariowa [77] experimented how different colours on rating scales effect how participants react to the colour coding in terms of mental effort and response time. The experiment showed that when more colours are used on a rating scale with numeric labels, the effectiveness of scales in acquiring user opinion also increases.

The literature discussing the effect of different rating scale designs on ratings of users suggest that there is a strong connection between the rating score obtained from a scale with how it is designed. When users receive strong visual and verbal cues from rating scale, it can allow them to analyse the rating process better, which can in turn affect how they are rating. The next section discusses the impact of mental efforts required in analysing rating scale design cues by users.

2.3 Use of Visual Heuristics for Rating

The previous sections discussed the research conducted in evaluating how different visual designs of rating scales result in difference in ratings from users. The difference in visual cues in rating scale designs elicit different emotional response in users, which results in the difference in ratings. Evaluation of products and ratings of products can considerably fluctuate due to the emotional state in which consumers are while rating them [95]. Marketing and advertising sectors have always attempted to appeal to the emotional cues of their targeted audience as a tool to increase and promote their brands. Similarly, in case of giving ratings to products using certain rating scales, users can interpret the visual and verbal cues of the scale and this can influence the ratings given on the scale. For instance, the rating scales which are fully labelled using verbal cues such as “very poor” to “excellent” or uses visual cues such as emojis triggering human emotions, can ignite emotional responses in users which can cause users to move away from giving extreme ratings [98]. The users are more thoughtful and engaged in rating process when using a rating scale that supplies visual and verbal cues to them [92]. This eventually helps them rate with different opinions they have for different products, instead of just randomly picking a response. However, according to [55], the polar emotions caused by the opposite emotional cues in the rating scales can also increase the cognitive load of users while rating using such scales. As a result of which, users tend to seek cognitive shortcuts, or Heuristics to reduce the time for decision-making while rating [93]. Heuristics can be defined as a mental shortcut that people take in order to to solve problems and make judgments quickly and efficiently. According to [93], when a rating scale uses strong verbal or visual cues that cause users to think more than they want to in rating process, they tend to use heuristics from the visual cues provided in the rating scale designs.

Two important hypotheses are proposed on how users interpret certain visual cues in rating scales in [92], **intensity hypothesis** and **familiarity hypothesis**. The familiarity hypothesis suggests that individuals become attached to visual and verbal cues with which they are familiar with and therefore, are more likely to prefer them. On the other hand, according to intensity hypothesis, more intense and strong visual cues can ignite intense emotional response in users [99], causing them to steer away from extreme response and be more cautious in how they rate a product. Both these hypotheses can help users find a heuristic or mental shortcut that can assist them in rating process. Visual heuristics used by users while rating has been demonstrated in more detail in the work by Tourangeau [90]. In this work, they extended the use of visual labels on rating scales from their prior work [91]. Their previous work [91] suggested that users have difficulty in using a rating scale (expressing their firm opinion) and thus, they rely on certain visual cues on the rating scale such as granularity, verbal labels or different colour hues at scale points by assigning meaning to these cues and using them in rating process. They proposed five separate heuristics that users might use while rating. Among the five heuristics, the last heuristic states as " Like (in appearance) means close (in meaning)"; which means visual cues that look alike have same meaning. This means when there is different colour or visual cues present in a scale (such as different colours at low and high ends), users will be more aware of the difference in different points of scale. The extended work by Tourangeau in [90] presented experiments which support the use of "Like means close" heuristic by users when using rating scales. The results suggested that using different labels and different hues at endpoints of scales can have definite impact on the rating behaviour of users as their rating can shift depending on the visual heuristics offered by the rating scales. Users tend to take more time and rate more positively, shifting away from low points, when different hues are used in endpoints. The majority of studies found that fine granularity (5-star scale), full labeling of scale points, presence of a neutral midpoint are considered to be the optimum and efficient design choices for rating scale. It has also been established that verbal cues and numeric labels can be effective in helping users in rating more effectively. However, the use of visual cues such as graphical icons, colours and shading are comparatively less explored avenues in this regard. A handful of studies exploring the effect of visual cues and colours in rating scales have provided findings which suggest that they can have strong emotional effect on users. When users are provided with apt visual cues which are designed to assist them in rating process, the credibility of rating increases while reducing the unwanted bias from rating scores. To understand the effect of using certain visual cues in rating scales, this research focuses the analysis based on the rating scale model used in Cena and Vernero [17]. In addition to visual cues, to further visualize how different colour metaphors can also be used by users as useful heuristics in rating process, this research adapted the traffic-light and warm-cool colour metaphor in rating scales from [14], [61] and [90] for designing the user study. The next section focuses on the consideration of user preference in rating scale designs by UX designers and how it can benefit the stakeholders in online marketplace.

2.4 User Preference in Rating Scale and It's Contribution

As mentioned earlier in many studies, one of the major concerns of online product reviews is whether or not the ratings truly reflect the opinions of users [9]. The role of eWOM and online feedback mechanism is unparalleled in online market and recommender systems. Without credible and consistent review from users, the recommender systems will fail at their core responsibility: to recommend users what they like. And for online business sphere, the user review is everything: rise and fall of a product or service solely depend on user's online ratings and reviews about it. Therefore, one of the most popular and important tool designed for collecting user's opinions: rating scale, certainly plays bigger role than the UX designers currently give it credit for.

The majority of studies which investigated the impact different design of rating scales on user's rating behaviour, have agreed upon the fact that allowing users to use a rating scale design they like, instead of a generic approach, can be far more effective for obtaining credible rating performance on online market. Allowing users to give ratings that align with their actual opinion is essential for the performance of recommender systems and online businesses such as Netflix or Amazon [76]. Maharani et al. [61] focused on user's perception in rating visualization and concluded that user's preference for rating scales should be taken into consideration as certain preferences in terms of rating scale designs can have an impact on user's ratings. This study asked users to rate using 5 different ratings scale designs (thumbs up/down, 10 points, 100 points, binary 5-star and 5-star) and then evaluated the choice of rating scales by users in terms of simplicity, ease of understanding, visual appeal and informativeness. The result showed that users have a clear preference in terms of using rating scale designs but it did not specify how this preference can affect rating scores. Another study with similar premise, conducted by Gena et al. [35], considered the effect of different rating scale design on rating scores. The study proposed to allow users to choose the rating scale design in order to promote user participation and satisfaction with recommender systems. The study showed that ratings given in different scale design for same item, then the ratings have lower mathematical correlation. This can result in inaccurate depiction of user opinion if rating scale designs are not customized based on. The study conducted by Cena and Vernero [17] suggested a set of guideline for designers for adjusting the rating scale options to fit users preferences, such as offering thumbs up/down for simple evaluations, whereas using star rating scale for general or default rating. The experiment Cena et al. [16] showed that the scores given on rating scales when users select them as their preferred scales are significantly different from ones they do not like.

In line with the user preference based model of rating scales by Cena and Vernero [17], this research proposes a user-preferred rating scale approach to investigate the effect of incorporating user's preference in rating scale scale in obtaining the true ratings. The growing literature on the impact of rating scale designs on rating scores and visual cues assisting the rating process have been the motivation for this research. The practice of allowing users to choose their own rating scale based on visual heuristics can be an important and useful addition to the UX design of online marketplace and recommender systems. The research goals are to understand contribution of visual icons and colour metaphors used in rating scores and preference of rating scales by user, and how this preferred scale design can be used to elicit true ratings from users. In order to validate the influence of user preference model and visual cues on the rating score of users, three hypotheses have been proposed. The research goals and motivation are briefly discussed in next (and last section) of this chapter. The next chapter encapsulates the hypothesis development and research goals for the thesis based on the literature review.

3 FRAMING RESEARCH QUESTIONS AND HYPOTHESES

The primary goal of product ratings is to reflect the product quality and it is very much essential in depicting product sales. The significance of user reviews and the biases related to it have already been discussed in section 2.1.3. In order to ensure credible user opinion acquisition, which is the core of online marketing and recommender system, the online rating scale design and their effect on user ratings are potential avenues of research. A number of factors while designing rating scale have been analysed for their role in helping users giving their honest ratings. Two theories can be considered in this regard: principle of non -redundancy and concept of satisficing. According to the principle of non redundancy in [36], the respondents tend to search for signals or cues on how to respond to survey questions. As a result of which, they tend to assign particular meanings to all incentives given in the format of question. According to concept of satisficing used in [50], it is also expected that the more demanding the 'cue-looking' / cognitively demanding a survey is, the more vulnerable a scale format is to response bias. These visual cues can be provided through changing rating scale design features, such as adding labelling to rating scales, either on all point or just end points, increasing or decreasing granularity of scales, adding or removing a middle neutral point in the scale, using different colours to label the scale or keeping the scale monochromatic. The effect of full labelling vs end labelling to reduce burden on users has been extensively analysed in [24], [97]. The effect of of putting a neutral midpoint in rating scale and the relative rating scale size has been evaluated in [92], which shows that presence of midpoint has helped reduce extreme response bias. The Effect of using both colour and labelling on endpoints of scale has been demonstrated in [90]. In this case, while the verbal and numerical labeling of the scale end points added minimal help to the participants, using different hues for the different ends of the scale proved to be a strong factor in influencing ratings from participants. Extensive research on various factors of rating scales (granularity, labelling, colour) has been done in [16], which shows sustainable evidence of scale granularity on differences in rating behaviour of users. Of all the rating scale design factors, the effect of using specific colour cues and special graphical icons have shown great potential, but it has not been extensively researched. Table table:research summarizes core findings of the works that have inspired my research goals.

Citation	Research Focus	Findings
Maharani, Widyantoro and Khodra (2016)	Rating Visualization, Rating Scale Preference, User Perception	Users prefer 5-star rating with different colours separating positive and negative ends, they prefer scales based on informativeness.
Cena and Vernero (2015)	Rating Scales, User Choices, User Interface	User preference for Rating Scale depends on informativeness, experience and visual appeal, 5-star is the most preferred scale.
Gena, Brogi, Cena and Vernero (2011)	User Preference, Rating Scales	User preferences for scales can cause difference in ratings, each rating scale can have their own "personality" based on their granularity and visual metaphor.
Toepoel, Vermeeren and Metin	Emojis, Likert-scale, Pictorial Answer Formats, Response Formats	Different rating formats produce different user opinions, smileys produced most positive user response.
Tourangeau, Couper and Conrad (2007)	Visual Features, Response Scale, Interpretive Heuristic	Using different colours at two ends of scales results in more positive responses from users.

Table 3.1: Research findings from important studies in rating scale visualization and user preference analysis

3.1 Research Context and Motivation

The goal of my research is to focus on user behaviour regarding presence and absence of strong visual elements or labels in rating scales and role of user preference in actual rating scores. It is already established that the users have certain reasons to prefer one rating scale design over another [61], [17] and the variety in rating scale design have immense effect on the ratings they give as well [18], [89], [90]. However, finding a relationship between the user preference in rating scale design and how that preference affect rating scores have not been explored. Similarly, the role of strong visual elements in both user's rating scale preference and the rating scores is a promising area which needs further research. Therefore, this research focuses on three factors and their interrelation: (1) strong visual elements in rating scales and (2) user's preference in choosing the rating scale design and (3) how both these factors influence actual rating scores and capture true ratings from users. Six rating scale designs are selected which have both neutral muted visual elements as well as strong visual cues that have proven to affect users emotional response while rating. The scales selected for the user study and research have the 5 point granularity, distinct neutral midpoint and no verbal/numeric labelling (since labelling and colour together already proved to make labelling redundant in [90]). The two design elements that are the modified in all six scales are colour cues and use of graphical icons. The choice of rating scale icons is based on the rating scale classification and model by Cena and Vernano [17], as well as from the findings in Maharani et al. [61]. The choice of colour cues on warm-cool and traffic-light metaphors is based on handful of of studies on colour coding including [90], [39] and [14]. Both factors are discussed in details in following subsections.

3.1.1 Use of Icons as Primary Visual Cues : Emoji and Star

Cena and Vernon presented a model for ratings in “A Study on User Preferential Choices about Rating Scales” [17] where they derived 13 features of rating scales from previous works and classified rating scales into 3 major clusters based on these features. The authors then conducted a user study with three rating scales, one from each cluster to rate two different objects. The user study attempts to figure out which scale design would the users choose to rate different objects and the motivation for users choice of scale. The classification of rating scales in this study is based on intrinsic and extrinsic features of rating scales deduced from previous works by [94], [35] and [69]. The three major classes of rating scales identified in [17] are:

- Human Scale: Scales which has a strong visual aspects which exploit human emotions to connote meaning for each rating point on scale. For this thesis, human face emoji is selected from Human scale.

- Neutral Scale: Scales which rely on quantitative inputs rather than visual characterizations. All the icons used in this scale are same and have no strong emotional/ visual connotations. For this thesis, star icon is selected from Neutral scale.

- Technical Scale: These are scales which have a strong focus on quantitative evaluations, similar to those of measurement tools. This class is omitted since sliders and other technical scales had worst review from users in [17] and [89].

Basing the user rating interface on the classification of rating by [17], I have used six rating scales : three from Human Scale class and 3 from Neutral Scale class. The Technical Scale class is excluded since the scales in this category are the least favoured by the users. Technical Scales have high granularity but no visual characteristics and require more effort on users part for rating task. From the user study and follow-up questionnaire conducted in [17] and [61], most users selected sliders (Technical Scale) as the worst scale and stated the lack of informativeness, visual appeal and ease of use, as reasons for doing so. On the other hand, most people who preferred star rating scale have selected having “right” granularity, familiarity and informativeness as the reason for doing so. Therefore, star and emoji are used as icons for 6 of the ratings scales (three of each) used in the user study and the motivations for doing so are as follows:

- Using emoji from Human Scale and Star from Neutral Scale as they have visual elements which Technical Scales lack.
- Using 5-points as the granularity for all the scales since this is the most popular and accepted granularity for most users.

Although the rating scale model and user study by Cena and Vernon [17], explored the influence of user preferences for a particular scale, but they did not consider how to use this rating scale preference can affect actual rating scores given on these scales. This is the area where the research model is built on. Previous studies by [36], [50] and [90] have already supported the tendency of users to look for visual cues while using rating scales. When rating scales include more visual cues, representing what each point on the scale means, users can interpret scales easily which affect their rating behaviour. The user rating scale model in this thesis is designed to explore the influence of visual cues on user's preference of rating scales and the effect it has on user's rating behaviour as well. In addition to graphical cues like star and emojis, colour metaphor on scales are also included. The role of colour metaphors in rating scales is discussed the the next subsection.

3.1.2 Use of Colour Metaphors as Secondary Visual Cues: Traffic Light and Warm-Cool Metaphor

Using colour hues to include more visual elements to help users have been researched in [90], [39] and [14]. Therefore I have added two different colour metaphors on each of my six rating scales (three of each) to analyse the role of colour cues as visual cues presented to users in rating scale as well. The cognitive impact of colour on heuristic consumer decision -making has been a widely researched material in both e-commerce and human computer interaction (HCI). For my rating scale model, I have chosen two colour cue models : *Traffic Light* and *Warm vs Cool*. Both the colour cues have been used in prior works to verify their effect on people and their judgements. However, the cognitive impact of color cues depends on the interpretation of what the colour cue "stands for". This can also be formally termed as the signaling significance [44] of colour cues about specific information that they carry in peoples' minds [62]. Individuals from different cultures can share a similar interpretation of the signaling value (meaning) of certain colours , such as red being negative or blue being "cool" colour or green standing for positive in spite of cross-cultural differences [75].

Warm vs Cool Colour Cue Model (Red Yellow Blue):

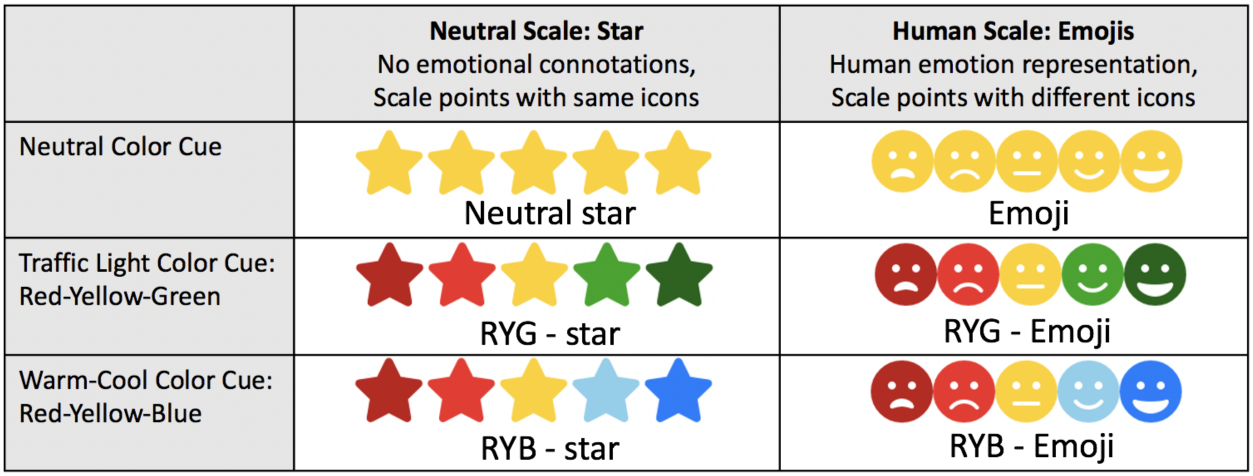
The concept of using warm colours such as red/orange to excite/arouse people than using cool colours such as blue/violet which has been experimented and validated in [20]. Similar colour palette has been used for rating scale in the work by Han et al. [39], to test the effect of red as a warm colour and clue as a cool colour as a heuristic for consumer decision-making.

The effect of using red as warm colour and blue as a cool colour has been applied as a visual cue for rating scales in the work by Tourangeau et al. [90]. The authors in [90] used rating scales with different colours for high and low end and same colour on both ends to compare how users interpret the scales and use them. The hypothesis in Tourangeau et al. [90] suggested that if two ends of a rating scale have different hues (shades of colour), then the respondents will also treat the endpoints as more significant than single coloured rating scales. This results in users putting in more effort in rating and reduced extreme rating. The different hues at different ends can also cause a positive shift in rating from users than single hues scales. Their experiment showed when the two ends of the scale had shades of a single hue (varying from dark blue to light blue), the users have interpreted the scales to have less significance. On the other hand, when the two ends were represented by shades of different hues (dark red at one end of the scale but dark blue at the other), the users could identify the low and high ends of rating scales more clearly and interpret the scale to have covered broader spectrum of choices. Similar colour cue as red-blue has been used in [61], which attempted to make a qualitative analysis of which one of the rating scale layout users liked more and why. Most users chose the 5-star scale with red in lo and blue in high ends. They chose this scale design due to its simplicity, readability, ease of use and informativeness. Based on the results from these two studies, I selected the warm-cool colour cue (red as warm colour, blue as cool colour) in my rating scales, both for star and emoji scales. The warm colour red is used for lower end of the scale and cool colour blue is used for higher end of the scale, since that is the norm for this colour cue. Out of 5 points on scale, the scale point 1 is the lowest so it has the darkest red hue, 2 has a lighter red hue. I have used yellow as the neutral midpoint of the scale. Light shade of blue is used to represent 4 and darkest blue is used to represent highest rating point 5 on the scales. The rating scales are shown in figure fig:Scalecue.

Traffic Light Colour Cue Model (Red Yellow Green):

Another very familiar colour metaphor used worldwide is the Traffic-Light colour cue: red for stop, yellow for pause and green for go. This colour metaphor is one of the most universally used visual cues for instruction around the world. As mentioned in [33], the dichotomy of green versus red is consistent across different color models and it is consistent across cultures (i.e. green="go" and red="stop"). Almost all countries of the world use red-yellow-green as their traffic light signals. For electronic appliances and devices too, red is used for stop button where as green as start button. Larrivee et al. [54] used 3 colour coded systems to classify food based on their nutrients into 3 color-coded categories: most healthy (green), medium healthy (yellow), or least healthy (red) based on calorie density and general nutritional quality. For rating scale design study, this colour metaphor has also been explored as colour cue in Bonaretti et al. [14]. Here, the 5-star ratings scales have been used to identify how using traffic light colour cues can trigger an emotional anchoring effect in ratings. The study suggested that using green colour can influence users to more positive ratings and using red can influence users for more negative ratings. On the other hand, scales which do not use different colours with specific meaning (i.e., monochromatic scales) do not have such emotional effect on users. As implied in [90], the use of different shades in rating scale has significant impact on rating behaviour of users, the traffic light metaphor can provide a comparative analysis with warm vs cool colour cue.

I have used traffic-light metaphor to see the differences between the two colour metaphors and how the users interpret each of them. The use of two colour also provide more options for users to chose their preferred scale design, which has not previously done in the literature. The scales used in the research and the visual cues assigned to them are fully illustrated in figure fig:Scalecue. One of the major motivations for choosing red as the negative/low end of rating scales for both colour metaphors is the polarity of the color red itself. The colour red has been reported to be associated with both positive (warmth, excitement) and negative (anger, danger) emotional responses [45] [84]. The 2020 work by Kawai et al. [47] analysed the polarity of red in two contexts: red alone and red with green. The research showed evidence that the colour red can be linked to negative responses when the colour green is presented with it, but not when red is presented alone. Their findings suggest that the colour red has a more neutral, even slightly positive effect without the presence of a contrasting colour pole. Only when green was present alongside red and a bipolar environment was presented to the users, the authors observed for the colour red facilitated responses to negative words were faster than to positive words. Basing on these findings from previous works, I used the colour red to present the negative/low end of the rating scale and blue/green to represent the positive/high end of the scale. Since people are more universally familiar with the Traffic-Light cue, comparing the effect of blue instead of green as a contrasting colour for red can give a different insight to how users will react to the scales.



	Neutral Scale: Star No emotional connotations, Scale points with same icons	**Human Scale: Emojis** Human emotion representation, Scale points with different icons
Neutral Color Cue	★★★★★ Neutral star	☹☹😐🙂😀 Emoji
Traffic Light Color Cue: Red-Yellow-Green	★★★★★ RYG - star	☹☹😐🙂😀 RYG - Emoji
Warm-Cool Color Cue: Red-Yellow-Blue	★★★★★ RYB - star	☹☹😐🙂😀 RYB - Emoji

Figure 3.1: Rating scales selected for study and the visual cues assigned to them

3.1.3 User Preference in Rating Scale Design

While most websites offer rating scale design of their own choice to online customers, the preference of users for rating scale designs can be very different from what is given to them. The importance of user's preferences in rating scale design was established in [17], where the authors had users choose between three scales, for rating two objects with opposite features. Results showed that user had different choices regarding the scale that they were given to use for rating. The findings showed that when users were given few options of rating scale designs to chose from, they choose different scales for different objects. The preference of rating scale design has also been the focus of [61], where users chose their favourite rating scales based on simplicity, ease of use and informativeness. Both these studies reached the conclusion that users can have different preference when choosing rating scale design for rating task. However, [61] just focused on the particular reasons why users preferred one scale over another and no actual rating task was involved. While [17] focused on the rating process as well, i.e., how users choose rating scales from different options when they have to actually rate an object using one of them. These findings clearly suggest that the preference of users in terms of rating scale design should be included in online rating process, but they do not explore whether the user preferred scale have any impact on the quality of ratings given by using them or not.

To the best of my knowledge, no research or study have been conducted to see whether including user preference for rating scales have an impact on getting the true opinions from users about the products. The novel aspect of this research is the investigation of user preferred rating scale model on rating score of users. The ratings scores on preferred ratings scale with other ratings scales used in the study are compared to differentiate rating scores given on them. Another novel approach taken in this research is the capture of "True" ratings from users. After the users are done rating the products with all 6 rating scales, I ask the users to select the "best" rating or ratings the product "deserved" from the 6 rating scores they gave for the product. Therefore, the acquisition of "True" user rating is not just data-driven or statistical, it is based on how users judge the rating scores they gave to a product themselves. This provides a stronger user-based validation in identifying the "True" rating scores from all the 6 ratings scores collected from 6 rating scales.

3.2 Research Questions and Hypothesis Development

The primary goals for this research are to analyse the effect of visual cues in rating scales on user ratings, to understand whether these visual cues can cause users to prefer one scale design over another and to investigate whether having a preferred scale design have any impact on the rating scores given on these scales.
Following the literature review of the influence of rating scale design on rating behaviours of users, I set out to explore the answers of the following research questions in mind:

1. Do users prefer the rating scale they are most accustomed to, or do they prefer scales which have more visual cues?

2. Do users rate differently on the scale that they prefer than on the scale they are most accustomed to? If so, how significant is the difference of rating scores they give using the two scales?
3. Do users give the rating that they truly believe the product deserves (called hereafter true rating) on a scale that they prefer?
4. Do users give their true ratings for a product using scales with more visual cues than scales which have less visual cues?
5. Does using rating scales which contain more visual cues impact the consistency of ratings given by users and cause any positive /negative shift of the ratings?
6. Does using rating scales which have visual cues impact the most frequent rating value given on a particular product?

In order to model the users' preference for rating scales and role of visual cues on rating score, I have proposed the following three hypothesis, which are correlated to each other

The first hypothesis is based on three previously explored theories: first, the principle of nonredundancy [36], which states that the respondents tend to assign meanings to visual elements/cues (colour, icons, space between points etc.) present in rating scales provided to them and use these meanings as a cognitive shortcuts or heuristics in the rating process. In other words, the persuasive power of visual elements such as human emoji or meaningful colour metaphor in rating scales can be applied to product ratings given by users as well. Second is the intensity hypothesis in [92], which states that when users are presented with meaningful labels/elements on rating scales which ignite emotional response in them, they tend to be more active and involved in rating process, since these emotional labels can increase their cognitive load during rating process. Third is the familiarity hypothesis [92], which states that users tend to prefer labels which are familiar to them at least to some degree. Therefore, the first hypothesis is based on the presence of both intense and familiar visual cues/labels in rating scales and which visual label/cue is more preferred by users while rating products with them.

H1: *Users prefer a rating scale which provides visual cues that assist users in their rating process, over rating scales they most familiar with on the internet.*

The second hypothesis is based on the user rating scale preference analysis in [89], [17] and [35], which show that user's preference in response formats have significant effect on ratings outcomes. These studies focus on separate factors in the design of rating scales and all of them showed that users prefer different rating scales and as a result, their rating scores can vary as well. If the first hypothesis can be validated, which shows that users already have different preference for different rating scale based on how familiar versus how visually informative a scale is, the second hypothesis is the second step in generating a user preference model for rating scale design.

H2: *Users provide their true ratings when using rating scale designs they prefer themselves.*

The work in [90] shows that using labelling and different shading on both ends of a response scale tends to push the responses towards the high (positive) end of scale, since the strong visual cues reduce the extreme negative responses in users. The hierarchy of verbal, visual and numeric label on rating scales have been analysed in [88]. This is the basis of the third hypothesis.

H3: *Using rating scales with more visual cues creates a positive shift and consistency in the ratings and true ratings are mostly given by users on these scales.*

Ratings scales have been initially used as tools to just collect user opinion for a product or service, while recent researches suggest they can do so much more. The hypotheses presented in my research and and their validations are aimed to investigate the role of rating scales as tools to assist users in giving their true ratings for a product. Certain visual cues such as colour and special icons can help users in making an informed decision while rating. The findings from the proposed hypotheses can provide insights for better UX design for rating scales in recommender/ review/ e-commerce systems.

4 RESEARCH METHODOLOGY

Rating system or rating scales have been widely used to capture the users' opinion on a variety of mediums online. The user rating of online products and services is an essential part of generating word-of-mouth which has a greater influence on purchase decisions than any other mediums [19]. With the advent of social media and social networks and new apps are flooding the screens that can do almost any task possible, users are involved in co-creating products and services [79] more than ever. Generating user-generated content largely depends on capturing user opinion and rating scale is one of the most widely used widget to capture user generated in a quantifiable manner. Rating scales are found in almost every website, app, social networking sites e-commerce sites and so on. Users do not just rate the products and services anymore, they rate their Uber driver, their employers, their teachers etc. These ratings are used in generating more personalized content for users online. As an important tool to capture user opinion and generating user-generated content, rating scale design and features can have a significant effect on the users rating behaviour i.e. how they rate. Effect of rating scale features on the user rating behaviour in recommender systems has been investigated in a number of literature. In most of them, significant differences in users' average ratings on different scales has been illustrated with substantial statistical evidence. Rating scales have been utilized as a persuasive tool in the research to elicit the accurate ratings from a user and to investigate difference in ratings based on the different use of colour and icons on the rating scales. How users interpret each scale and how this interpretation eventually influences the rating score on each scale have been the motivation of this research. An interactive interface has been used to collect the user ratings by conducting a user study with active participation and it received behavioural ethics approval with approval number BEH# 1521 on December 5 2019. This chapter encloses the research goals, methods, motivations and user data collection design for the research.

4.1 Research Model

The research aims to investigate the following: role of visual cues and colour metaphor in user preference of rating scale, getting true ratings from users and effect of visual cues and colour metaphor in actual rating scores. The three hypotheses proposed in previous chapter deal with all three of the above research goals respectively. For collecting data, I have designed a user interface which asks user to rate some day to day products using six different rating scales. These six scales differ in two major visual features: colour and icons. I have used three different colour schemes to differentiate low and high ratings on a scale. The traffic light metaphor (red-yellow-green) and warm -cool metaphor (red-yellow-blue) and for neutral approach, I have used the common yellow colour mostly used on rating scale. For icons, I have chosen neutral star icon with no strong connotations and standard human emojis (smileys) which has strong visual metaphors to express low score with sad face to high score with smiley face. The granularity of each scale is 5-point which is favoured by users most according to [61]. The six rating scales used for the user survey have been selected based on the rating scale model provided by Cena and Vernano [17]. The two colour schemes/cues selected have also been based on prior works which are discussed in the previous chapter. All rating scales are shown in figure fig:ratingscale. The rating scales are named as follows:

- All yellow stars: Common-star scale.
- Red yellow and green star where red connotes negative, yellow neutral and green as positive values on the scale: Red-Yellow-Green star or ryg-star scale
- Red yellow and blue star where red connotes negative, yellow neutral and blue as positive values on the scale: Red-Yellow-Blue star or ryb-star scale
- All yellow emoji: Emoji scale
- Red yellow and green emoji where red connotes negative, yellow neutral and green as positive values on the scale: Red-Yellow-Green emoji or ryg-emoji scale.
- Red yellow and blue emoji where red connotes negative, yellow neutral and blue as positive values on the scale: Red-Yellow-blue emoji or ryb-emoji scale.

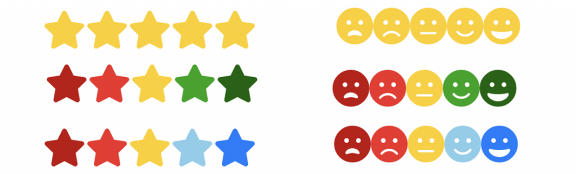


Figure 4.1: Six different rating scales chosen for the user study

The research process is divided into five stages. The block diagram for the research model is shown in figure fig:model.

Design, Development and Deployment of User Study Interface: The first stage involves designing a user study interface to collect data on different rating behaviour for different rating scale. The user study is designed to collect the data in three different phases which is discussed in next section.

Data Processing and Mining for Hypotheses Analysis: In order to use the data for hypothesis analysis, the data from all three phases of user study are collected and categorized to generate usable CSV files. The data collected are : ratings scores of each product using each rating scales, preferred rating scale of each user, most commonly used rating scale by each user and true rating score of each product by each user.

Validating H1 with Boxplot, Histogram and Correlation Analysis: First hypothesis is validated by using distribution of actual ratings given on rating scales which users preferred and the ratings given on rating scales which users find common. The direct relation of ratings given on these two user selected scales are shown using Spearman correlation analysis.

Validating H2 with Frequent Pattern Mining: Second hypothesis is supported by using frequent pattern mining technique called Apriori algorithm to find the most frequently selected rating scales to give true ratings for a product.

Validating H3 with Mean , Mode and Median Analysis and Wilcoxxon Test: The third hypothesis is supported by performing mean. median and mode analysis of rating scores given on each rating scales by the users. The resulting differences in ratings cores are validated by Wilcoxxon Ranks test.

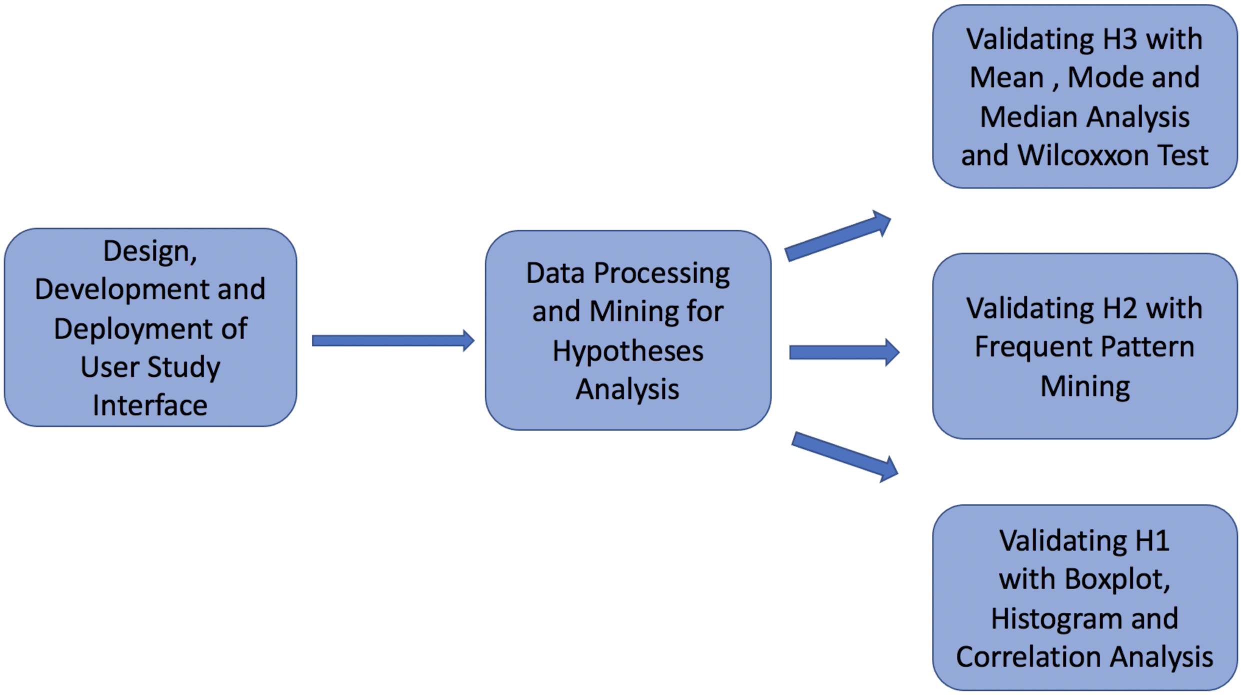

Design, Development and Deployment of User Study Interface
Data Processing and Mining for Hypotheses Analysis
Validating H3 with Mean , Mode and Median Analysis and Wilcoxxon Test
Validating H2 with Frequent Pattern Mining
Validating H1 with Boxplot, Histogram and Correlation Analysis

Figure 4.2: Block diagram of the research process

4.2 User Study Design and Implementation

In order to validate the hypotheses, I designed a web-based user interface with the six rating scale designs mentioned before. The user study includes a list of 21 day to day products that users are familiar with. Users have to select which products they have used before and rate those products using each of the six rating scales. In order to collect user preference of rating scale and true rating scores, I decided to ask users by a survey to select the most preferred rating scale design, along with the most commonly seen rating scale design among the six given scales after they are done rating. The true ratings of users, i.e., the rating score that a user thinks a product deserve is also collected from the users which is a novel approach in the user study for rating scale. The user study design and survey process are discussed in next section.

4.2.1 User Study Interface Design

The user study interface is developed using “React”. React is an open source, declarative, front end JavaScript library used for building interactive user interfaces. The interface is deployed and managed on Heroku: a container-based cloud Platform as a Service (PaaS). The survey application follows a 3 -tiered architecture that contains:

- A back end written in Node.js which is an open source server run time technology that runs in Javascript and it is deployed in Heroku.
    - Serving front end when users browse it.
    - Hosting API for front end to communicate via RESTful requests.
    - Processing request and communicate with data storage component.
- A front end written in React.js which is an open source web application technology.
- A data storage service, in this case I have chosen Google spreadsheets. The back end operates on CRUD (create, read, update and delete) on the data via google APIs.

The user survey web application itself walks the participants through a series of pages, each containing certain questions and survey, collects the answers along the way as a JSON object under the hood. Then it sends the JSON objects to back end which processes the objects and adds to spreadsheets under my google account. The Data can be collected as CSV formats from the google drive.

4.2.2 User Study Implementation

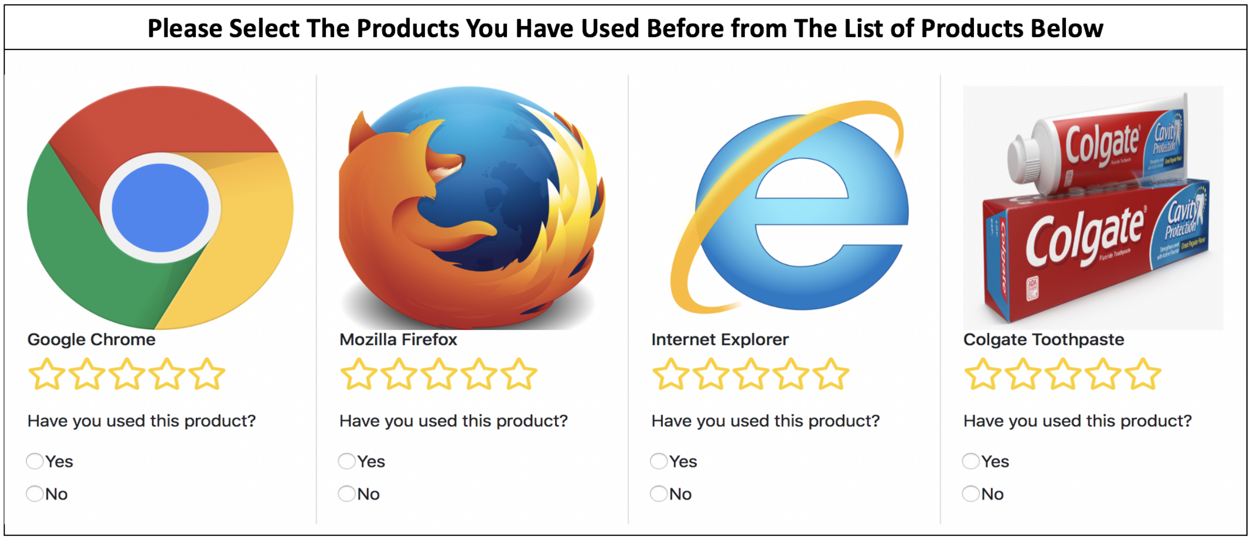
The user rating scores and rating scale information are collected in three consecutive phases from the user study. At the beginning, it asks the users for their consent for participation and informs them about the anonymous nature of data collection of the survey. Then users are shown a list of 21 common products and asked to select which products they have used before i.e. have experience with. After users have selected their products, they are asked to rate each of the selected product using all the six ratings scales (figure fig:ratingscale).

Once the rating process is completed for all the selected products, users are asked to select the scale which they would prefer most to rate the products and the scale which they have seen mostly on internet. At the final phase, the users are shown the numeric value of their ratings given on each scale, such as 3, 4, etc. on a scale of 5 for a product and they are asked to select the number they think is most deserving for the product.

This is repeated for every product they have rated in the user study. The whole process can be classified into the following three phases and data from each phase is used to validate the proposed hypotheses. The three phases of user data collection are further described below.

Phase 1: Select Products and Rate Them with All Six Scales

The primary idea for choosing items to rate was to include items which users are familiar with. I designed my primary study with movies to rate. However, most users for the pilot study hesitated to rate movies since a lot of them were not familiar with the movies I listed for them to rate. From the feedback of pilot study which included 14 users, I changed my items to common products. The users from pilot study also suggested that only rating movies was getting monotonous for them. In order to keep the users from getting used to rating the same type of products, I selected 7 specific item types and included 3 brands for each item type. The list of the products is given in table table:product. Since all products are chosen from different areas, ranging from technology to daily usage such as toothpaste, to common fast food chains, users have a variety of products to rate from. This helps them from repeating the same ratings for all items. At the same time, since users are asked to only rate products they have used before, their ratings is based on their real experience than just random ratings they picked to finish the survey. In the first page, the users are asked to select the products that they have used displayed on the page by clicking yes/no. Right below the products, the users will be asked to rate the product using the neutral-star scale as shown in figure fig:productchoice. After the users have selected the products and rated them once, they will be shown their selected products individually with one of the rest 5 rating scale below and asked to rate the products again. Both the products and rating scales appears in random order to the user as shown in figure fig:productrate so that the user does not get much opportunity to remember exactly what rating they gave to a particular product before.



Please Select The Products You Have Used Before from The List of Products Below

Google Chrome
☆☆☆☆☆
Have you used this product?
○Yes
○No

Mozilla Firefox
☆☆☆☆☆
Have you used this product?
○Yes
○No

Internet Explorer
☆☆☆☆☆
Have you used this product?
○Yes
○No

Colgate Toothpaste
☆☆☆☆☆
Have you used this product?
○Yes
○No

Figure 4.3: First page of user study: Users select and rate products of their choosing

Item Type	Item Brand
Soft Drinks	Pepsi RedBull Sprite
Email Provider	Gmail Office 365 Yahoo! Mail
Internet Browser	Google Chrome Mozilla Firefox Internet Explorer
Photo Editing App	Adobe Photoshop Picassa Microsoft Photos
Social Networking sites	Facebook Twitter LinkedIn
Toothpaste	Crest Colgate Sensodyne
Fast Food Chain	KFC Pizza Hut Domino's Pizza

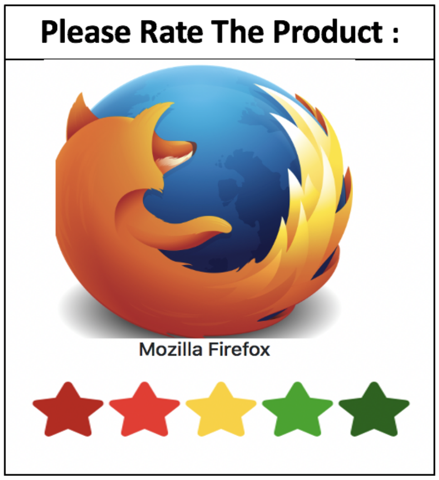
Table 4.1: All the products used in the user study and their groups

Phase 2: Select Preferred and Most Common Rating Scales

After all the user selected products are rated using all 6 different scales, the users are shown all 6 scales and asked to select the scale that they would prefer using if they have to rate the products again, shown in figure fig:preferredscale. Next, they are shown the 6 scales again and asked which one of these scales has the user seen mostly around internet.

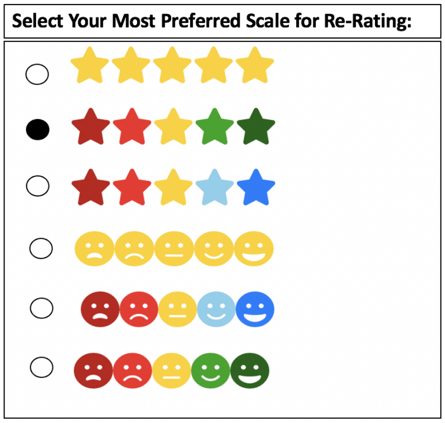
Phase 3: Select Rating Score Most Deserved by The Product (chosen rating)

The final phase of user study shows each user the products that they have rated again and below the product, they are shown the numeric value of the rating (1/2/3/4/5 out of 5) they have given to the product on each of the six different scales. The important thing to note here is that I have only shown the numeric value of the ratings provided by users here, not the scales on which they rated, as presented in figure fig:numeric. For



Please Rate The Product :

Figure 4.4: Selected products appear with one of the rating scales in random order



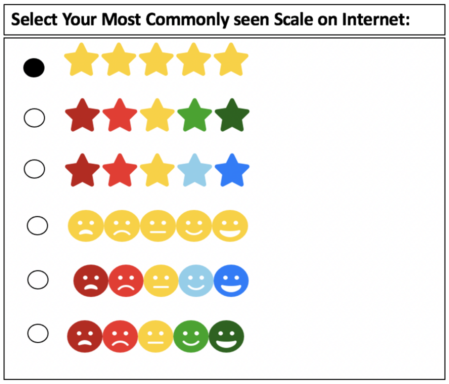
Select Your Most Preferred Scale for Re-Rating:

Figure 4.5: Users asked which scale they prefer to use to rate products again

Mozilla firefox, the user have given 2 out of 5 on yellow common star scale, 4 out of 5 on RYG-star scale, 5 out of 5 on RYG-Emoji scale, 4 out of 5 on RYB-Emoji scale, 3 out of 5 on Emoji scale and 3 out 5 on RYB-star scale. But here, only the numeric value of the rating score is shown to user. Users are asked to select the rating value they think the product deserves. In this case, the user thinks 3 out 5 is the correct rating for Mozilla Firefox, so he/she has selected 3 out of 5. If a user has given the same score on multiple scales, selecting only one will automatically select the same values. Both the score and the scale on which the score is given are stored from the page. For example, in figure fig:numeric, the user selected 3 out 5 as the chosen rating, so the value 3 and scales Emoji and RYB-star are stored as data from this page.

4.2.3 Data Collection and Participants

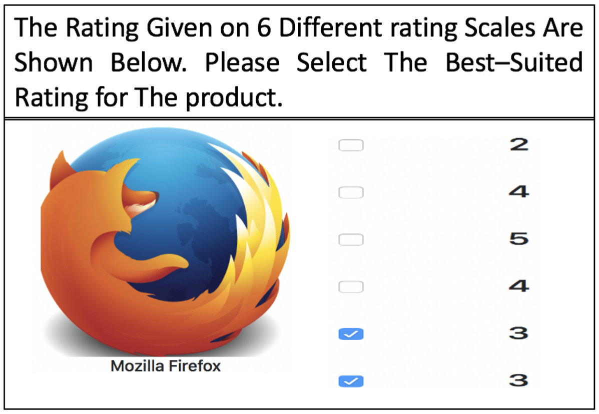
Participants for this study were mainly collected from undergraduate/graduate students of the University of Saskatchewan taking courses in Summer and Spring 2020. The setting for user study was primarily PAWS:



Select Your Most Commonly seen Scale on Internet:

●
○
○
○
○
○

Figure 4.6: Users asked to choose the rating scale they mostly see around internet



The Rating Given on 6 Different rating Scales Are Shown Below. Please Select The Best–Suited Rating for The product.

□ 2
□ 4
□ 5
□ 4
☑ 3
☑ 3

Figure 4.7: The numeric value of ratings given by users on six rating scales are shown to users

Personalized Access to Web Services which is the online platform for students of University of Saskatchewan. Recruitment for the study was done through announcements by me and my supervisor on university website and the sending of the survey link to the students via facebook, LinkedIn and PAWS. All the participants were at least 16 years old. Before the main study, a pilot study was conducted to test the validity of my user study tools and methods. For the pilot launch, I previously used set of 18 movies, ranked into highly popular, average and less popular categories (6 of each) to rate with same rating scales. For the pilot study, I recruited 14 random students from the university. The pilot study revealed that the participants were mostly not satisfied with the list of movies and they were hesitant to rate movies they haven't watched before. Even if they had the choice of selecting movies they have watched and only rate them, majority of participants were unfamiliar with the a lot of the movies I used. After the pilot study, I changed my rating items into day to day products with a variety of product types for the users to rate.

A total of 203 completed responses were received. The study generated random id for each user so the identity of the participants was completely anonymous. While pre-processing the data, some outliers were discovered with empty response files and duplicate responses from same id. After removing the outliers, final tally of valid responses was 187. The users were asked about their gender, age and nationality and the demographic of the users are given in table:demograph. The users were mostly young adults and students with working knowledge of the online rating systems.

Total participants =187	

Gender	Male=79, 42.2%
	Female=108, 57.7%

Age	16-24=98, 52.4% 24-35=67, 35.8% 36-45=13, 6.9% 46-55=9, 4.81%

Table 4.2: Demographics of all participants

4.3 Research Framework and Application

The research explores the effect of including user preference in rating scale design and meaningful visual cues in rating scales. The end goal of this research is to determine the significance of including both user preference and visual cues in rating scale designs in online platforms.

Figure fig:frame shows the contribution of all three hypotheses in developing an improved rating scale model. In the figure, the "User Preference Based Rating Scale Model" symbolizes a proposed rating scale model that allows users to use their own chosen rating scale design and includes different visual cues in rating scales provided to the users to choose from. The boxes A and B contain factors need to be included in the rating scale model while C, D and E contains the findings from three hypotheses that support the inclusion of A and B in the model. H1, H2 and H3 are the three hypotheses proposed in chapter 3. The first hypothesis or H1 establishes that user's preference of rating scale design is usually different from the rating scale they see on the internet. therefore, confirming H1 establishes the requirement of adding user-preferred rating scale (C) in the proposed rating scale approach. Now, the next hypothesis is designed to see how the preferred scale performs in terms of getting true ratings from users.

The second hypothesis establishes that the scale preferred by a user is more likely to get true ratings from the user for a product than other scales (D). H1 already supports the fact that users prefer different scale designs than the one they are used to seeing and H2 supports that the preferred scale is the one users want to give their true ratings on. The confirmation of both H1 and H2 therefore can be used to support inclusion of user preference in rating scale (A). The results of H2 is based on the user's own opinion about which rating score they think is the best for a product, collected in phase 3 of user study.

The third hypothesis (H3) investigates how the visual cues affect the actual rating scores given on the scales. The confirmation of H3 shows that users give more consistent and positive ratings (E) on scales that

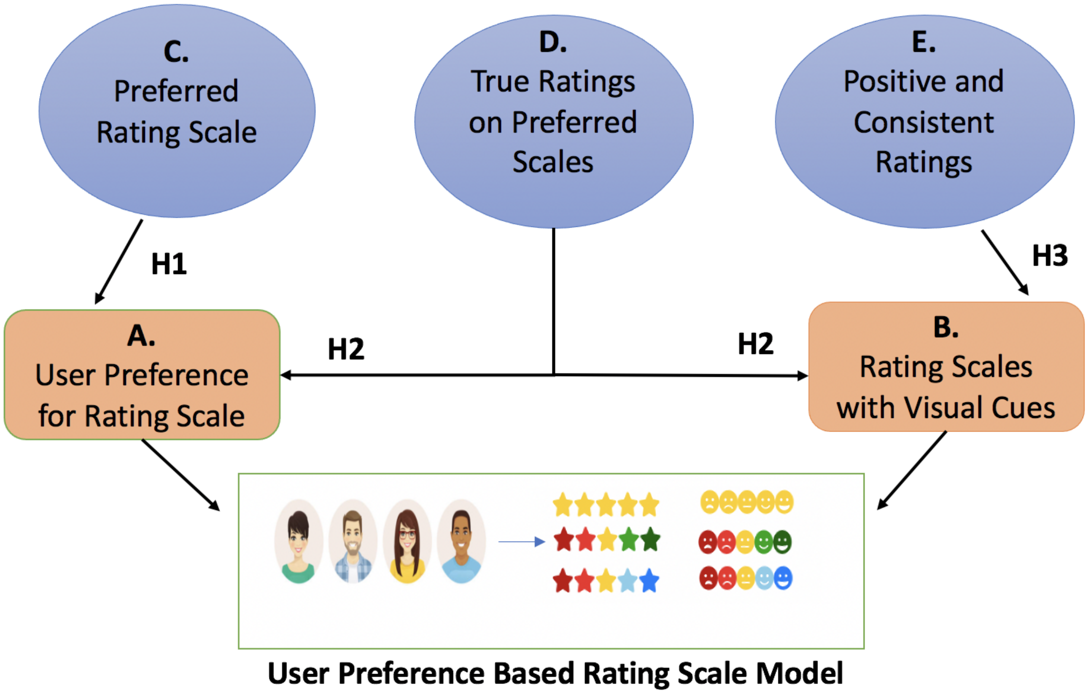

C.
Preferred Rating Scale
D.
True Ratings on Preferred Scales
E.
Positive and Consistent Ratings
H1
H3
A.
User Preference for Rating Scale
H2
H2
B.
Rating Scales with Visual Cues
User Preference Based Rating Scale Model

Figure 4.8: Conceptual framework of User Preferred Rating System Model based on findings from proposed hypotheses

contain more visual cues. In addition, H3 also shows that users mostly use rating scales with more visual cues to give their true ratings. The confirmation of H2 states that true ratings are given on preferred scale and confirmation of H3 asserts that the true ratings are given on scales which have more visual cues. Therefore, H2 and H3 both can be used to support the use of visual cues (B) in rating scales to persuade users to give more positive as well as true ratings for a product. Therefore, since the factors visual cues (B) and user preference (A) both produce true ratings (D) and positive ratings (E) for online commercial sites, these factors are worth being considered to be included for improving current rating scale systems.

5 USER PREFERENCE ANALYSIS

This chapter describes the validation of first hypothesis, which investigates the preference of users in rating scales and how the presence of strong visual labels in scales can affect the preference. Prior researches stated in previous chapters have investigated and stated the fact that reviewers take help or cues from certain visual elements or labels presented in the rating tools and interpret them to make rating decisions. This can affect the numeric evaluations collected from online review/recommender/e-commerce systems. Understanding how the users interpret certain UI elements and visual cues and how certain visual element influence their rating decisions can help the UX designers to make the right choices and strategies when it comes to rating tools. Letting users choose the visual cues and elements such as icons used in rating scale or use of colour, can be one of the UX decisions which help the rating scores given on such scales reflect the true opinion of users. This different preference of users for different rating scale designs can translate into how they rate an item using a scale.

5.1 Preferred Vs Most Common Rating Scale Designs

When users are asked to rate an item, they tend to use certain visual elements of the rating scale such as colour, granularity, graphical icons to assign meaning to the rating scale points. These interpretations often act as a cognitive shortcut or heuristic for users to be used in product rating [90]. For example, in [90], when different colours or verbal labels are used in two ends of scale, users use this as a heuristic to visualize difference between the end points of rating scale to be more intense than scales which use only one shade. This interpretation is reflected on their ratings as well and they rated more positively with such scales than with scales which are monotonous and had no label.

My first hypothesis is based on how user's rating scale preferences depend on the visual cues and familiarity of rating scales while rating the products with them. We asked users to rate a list of items with 6 different rating scales with different visual cues and neutral elements. The users were asked to select their most preferred and most familiar rating scale designs among the scales. The performance of each scale in terms of user's preference and familiarity as well well how ratings given on both selected scales for each user is used to validate the hypothesis.

5.1.1 Statistical Analysis and Evaluation

The data collected in first, second and third phases have been used to carry out the empirical evaluation of the following findings:

Hypothesis 1: Users prefer a rating scale which provides visual cues that assist users in their rating process, overrating scales they most familiar with on the internet

Design: Four factors (user's preferred scale, most common scale, rating scores given on preferred scales, rating scores given on most commonly seen scales), within subjects design.

Subjects: Anonymous participants (187) were asked to rate their used products with all 6 rating scales and select their most preferred and most commonly seen rating scales among them.

Apparatus: The raw data was collected through the user interface web-application and stored on google spreadsheets of the researches account via Google APIs.

For the empirical evaluation , both descriptive statistics (histograms) and inferential statistics (Wilcoxon signed-ranks test and Spearman correlation) have been used on the rating data for analysis.

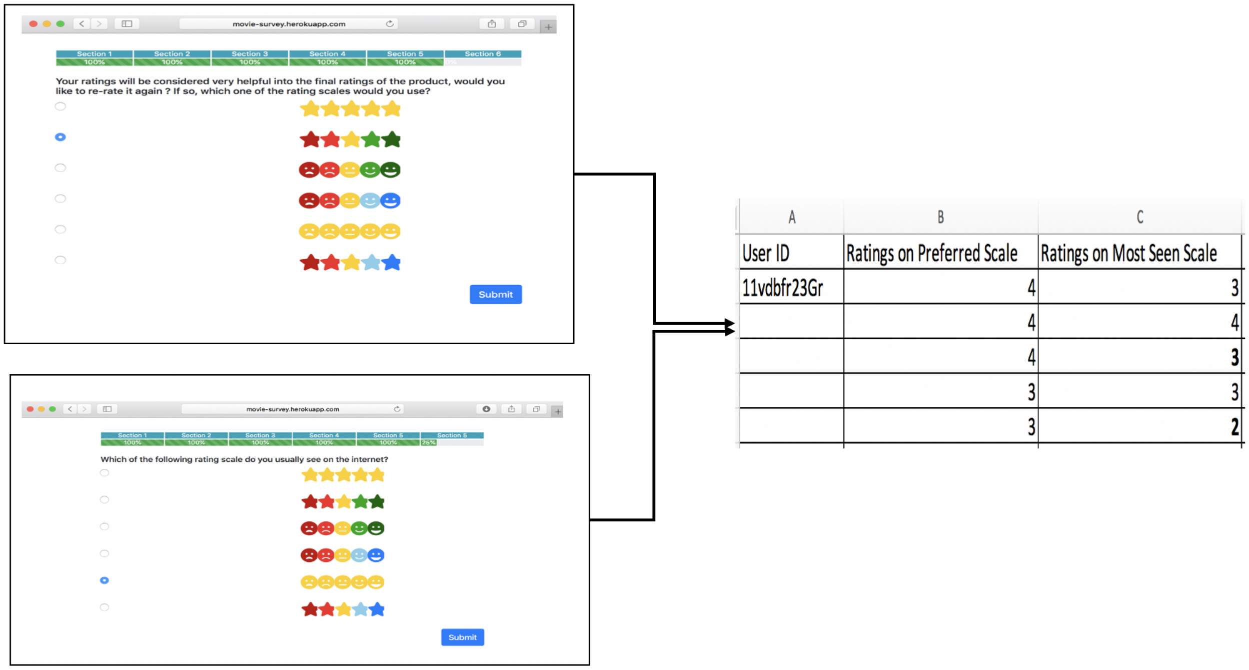
Data Collection

The users were asked to select their most preferred rating scales among the six scales and the scale they are most familiar with form the six rating scales. When each user selected their choice of most preferred and most familiar scales, the collective ratings for each item they rates specifically using these scales were retrieved. For example, suppose a user selected RYG-Star as his/her preferred scale and Emoji as the most familiar/most seen scale. In that case, the ratings that the user gave for every product using RYG-Star scale is collected as ratings on most preferred scale and the ratings given for each product using Emoji scale by the user is collected as ratings on most seen scale, shown in figure file.

Selection of Rating Scales by Users

Figure preferred vs common shows how often each scale was selected by users as their preferred and as the most seen /familiar scale of all 6 scales. The most preferred scale was the RYG-Emoji scale and the most commonly seen scale was the Neutral scale. Figure preferred vs common also depicts the choice of users in terms of rating scale designs and it is a clear indication that users have mostly preferred a different rating scale design as their preferred one rather than the one they commonly see around the internet. Most users (54.75%) have chosen RYG-Emoji as their preferred scale design, whereas the largest population of 86.5% users selected the Neutral-star as their most commonly seen rating scale design. It can also be noted that only 2.2% have selected the RYG-Emoji as commonly seen scale design. This is a clear indication of user's preference of rating scale design being very different from the rating scale they have used or seen before.

To understand how each user's rating behaviour changed according to which rating scale they are using,



Your ratings will be considered very helpful into the final ratings of the product, would you like to re-rate it again ? If so, which one of the rating scales would you use?

Which of the following rating scale do you usually see on the internet?

A	B	C
User ID	Ratings on Preferred Scale	Ratings on Most Seen Scale
11vdbfr23Gr	4	3
	4	4
	4	3
	3	3
	3	2

Figure 5.1: Data for preferred and common scale selected by users

the boxplot and histogram of all the ratings that users have given in the preferred scale and most commonly used scale for the products are shown in figures fig:box and fig:hist.

The visualization of rating scores on both user selected scales show clear differences between the ratings given on them and none of them follow normal distribution. When users are rating by using the scale they most prefer, product ratings seems to be notably different from the ratings given by using the most seen scales. For further statistical evaluation, Wilcoxon signed rank test is conducted to determine if the differences between rating scores given on two user selected scales are statistically significant. Wilcoxon signed rank test is a non-parametric version of the one-way ANOVA, is selected for this analysis, since I have two independent variable groups (preferred Scale and most seen scale) and the dependent variable values(rating scores on each scale) are ordinal data. In order to understand the relationship between the overall ratings provided on the most preferred and most commonly seen scales, correlation analysis is also carried out for both of them. Since the data does not follow a normal distribution, Spearman correlation analysis was chosen.

5.1.2 Result and Interpretation for Hypothesis 1

The goal of the first hypothesis is to establish the notion that a rating scale's rating outcomes/scores vary with the level of familiarity and level of fondness/preference users have for a certain rating scale design. The presence of strong visual elements can be a factor for the preference of rating scales in users. The presence of such visual cues help users in rating process [61], [90], therefore users prefer such visually rich scales than scale designs they are familiar with. The ratings given on users on preferred scale is different than the ratings given on familiar scale, which indicates that users rate differently when it comes to rating scales they like

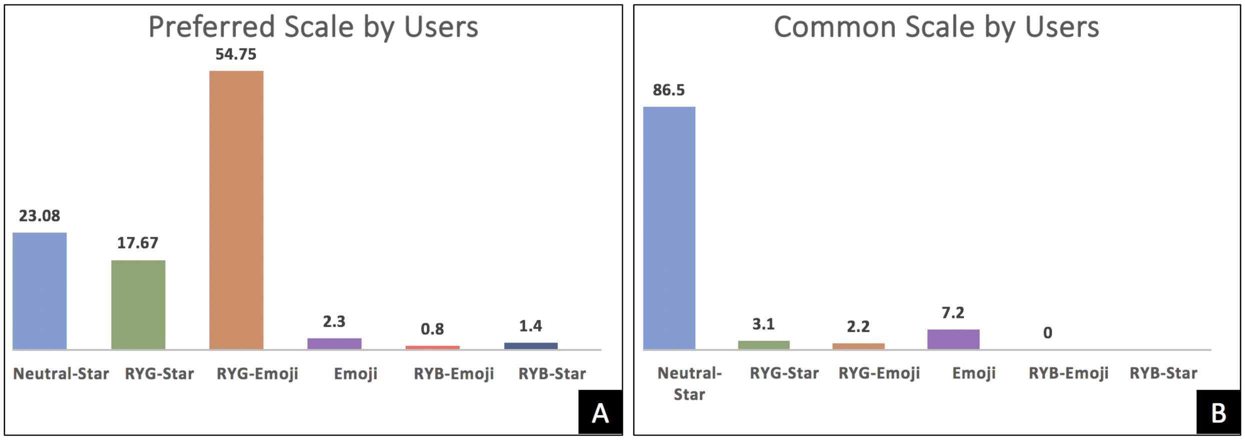


Figure 5.2: Frequency of choosing scales as (A) most preferred and (B) most seen

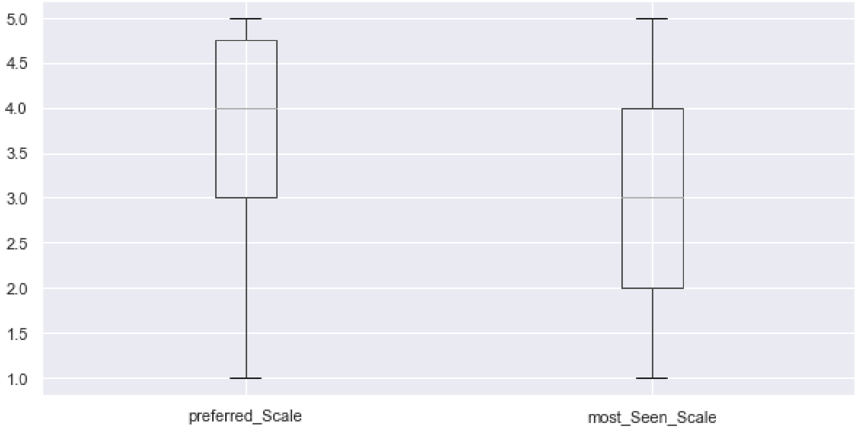


Figure 5.3: Boxplot of Preferred vs Most Common Rating Scale Ratings

than the one they normally use.

Comparison of Rating Scores on both Preferred and Most common Rating Scale

For the within-group experimental analysis, I collected the rating of each user given on each product they selected on their preferred and most commonly seen scale. For instance, if a user has selected RYG-Star as his/her preferred scale, and Emoji as the most common/seen scale, then the ratings given on RYG-Star scale and the ratings on Emoji scale for all the products he/she has rated is collected for data analysis (see figure file). Figure preferred vs common shows the frequency of users selecting preferred scale and most common scale out of the 6 rating scales. Overall, most users have preferred different scales than the one they normally see online. The result reveals that 54.75% users have selected RYG-Emoji as their preferred scale, where the visual cues are provided both in terms of icons(emoji) and colour (Traffic-light metaphor). Therefore, it can be stated that users prefer rating scales with more visual cues which can assist them in rating process. The most preferred scale RYG-Emoji is not a commonly seen scale since only 2.2% users have selected this scale

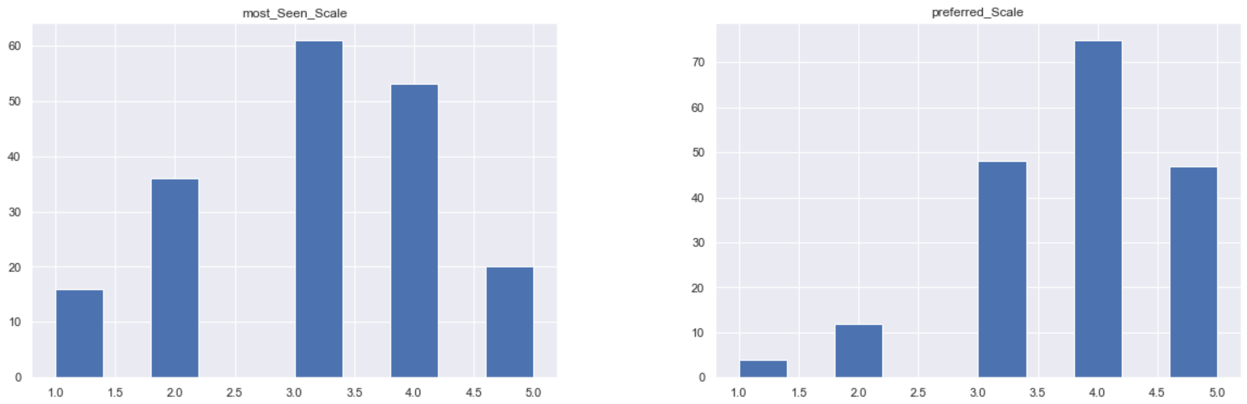


Figure 5.4: Histogram of Preferred vs Most Common Rating Scale ratings

as commonly seen scale. This states that users selected a visually rich scale (RYG-Emoji), which is not much familiar to them as their most preferred scale.

On the other hand, 86.4% users selected Neutral-Star as the most commonly seen scale. But, a rather low percentage of 21.08% users have selected Neutral-Star as the preferred scale. Another important takeaway from this analysis is that the Emoji, RYB-Star and RYB-Emoji are the least preferred scale for all users. Since the red yellow blue or warm-cool metaphor is not as widely recognized as the traffic light metaphor, the colour scheme in this case could have favoured the more widely known traffic-light metaphor.

Statistical Difference between Preferred and Most Common Rating Scales

Table 5.1: The correlation coefficient of ratings on preferred and most seen rating scales

	Preferred Scale	Most Seen Scale
Preferred Scale	1.00000	0.249519
Most seen Scale	0.249519	1.00000

The Wilcoxon signed-ranks test is conducted to see if the difference between the rating scores given on most preferred and most common scale is statistically significant or not. Since the data contains two independent groups (the two rating scales preferred and common) , Wilcoxon test is selected as the proper fit for such analysis[63]. The analysis shows significant differences between the ratings given on most preferred scale and ratings on on most common/seen scale by each user ($p<0.005$) shown in table 5.2. This validates the difference in product ratings given by users when they are using preferred scale versus using the scale they are more familiar with. In order to investigate and visualize the actual nature of this difference, or the nature of the relationship between most preferred and most seen scales, I calculated the Spearman correlation between rating scores of two scales. The correlation analysis yields the results in table 5.1. The results show that the ratings users gave on their own preferred scale and most seen scales have a weak correlation. Following the Cohen's classification system [22], only the largest relationships i.e. where the correlation coefficient r >0.5

have been considered to be significantly correlated. Following Cohen's classification, the relation between the ratings on these two user-selected scales are not strong enough to have any significant effect on each other. This means the increase or decrease of ratings in preferred and most common scale do not depend on each other. The users tend to give different ratings using both of them. This finding has been used as the basis of my first hypothesis.

This difference is finally visualized by using histogram and boxplot for both rating scale groups as shown in figure fig:hist. The boxplot graph shows clear difference in the central tendency of rating scores on both scales, where the rating scores on most seen scales tend to have e wider spread compared to preferred scale fig:box. This can be interpreted as users giving inconsistent ratings, on wide range of values when they are using the most seen/common rating scale. On the histogram, the distribution of both rating scales are shown where the preferred scale ratings seem to have more skewed distribution than most seen scale and has a shift towards positive ends. This further establishes that user ratings are much more consistent and positive when using preferred scale.

Finally, results from Wilcoxon test shown in table 5.2 validates the strong difference between ratings provided by users in both types of rating scales. The statistical and empirical evaluations support the first hypothesis which claims that users tend to prefer rating scales which provide them visual cues as this can ease their decision making process while rating. The findings also revealed that ratings on preferred scale tend to shift toward more higher end than most seen/common scale. Including rating design options for users to choose from can make a significant difference in rating scores. The impact of preferred scale and visual elements in rating scores and their analysis are presented in next chapters.

Table 5.2: Wilcoxon signed rank test for ratings on preferred and common scale on all products

Pairwise Comparison	**Z-Value**	**p-Value**
Preferred Scale and Most Common Scale	-5.58	**0.000**

6 EFFECT OF USER PREFERENCE IN TRUE RATINGS

In interactive online systems today, users are primary contributors to the systems themselves. In many such consumer-driven systems, the users are given scope to customize and modify the user model and add their own preferred settings to help them navigate through the system better [18]. Online shopping websites such as Aamazon or ebay set recommendation for users based on their shopping or search preferences, targeted ads are generated based on how users are using the search engines. In order to collect user's online shopping preferences, these systems use rating scales with different features such as granularity or visual presentation. For example, Amazon uses star ratings, Facebook uses thumbs with emojis, YouTube and Netflix use thumbs, Tripadvisor uses circle and so on. Researches already show that users have strong preferences when it comes to choosing rating scale designs based on how informative, visually helpful they are [17], [61], [88]. But, user's preferences regarding rating scales and the effect of such preference on the overall rating scores have largely been untapped.

The first hypothesis established that users prefer rating scale designs which have visual elements and they are often different from the ones they see online. The next hypothesis asks the question: can this preference of rating scale design bring out true ratings from users more, compared to other scales? Prior studies in this area collected rating scores from rating tasks given to users in surveys. In my user study, after users are done rating products with all six scales, I asked the users which of the six rating scores given on a product they themselves *believe* is best suited for that product. In this case, the "true" ratings are ratings that users chose themselves after rating tasks are completed. Therefore, my first hypothesis is tested with respect to data which that users themselves have approved after providing the ratings. This is termed as "user-approved" ratings hereafter. This is a novel approach in terms of user data collection and for deriving the best-suited/true ratings from users. The results of the user study will decide whether the scale preferred by users reflect the true ratings of users (selected by users themselves) or not, based on users' own opinions about what true ratings are. The findings from this experiment can be used to compare the quantifiable impact each visual cue has on the rating score in later sections, which is the base of my third hypothesis.

6.1 Data Processing and Evaluation Procedure for Second Hypothesis

The user study has been divided into three phases : users are asked to select the products they have used and rate them with 6 rating scales in a random fashion, then they are asked to select their preferred scale and the scale they see most on the internet or "common" scale and then they were asked to select the numeric value of the ratings they provided for each product.

The data collected in second and third phase have been used to carry out the empirical evaluation of the following findings:

Hypothesis 2: Users provide their true ratings when using rating scale designs they prefer themselves.

Design: Four factors (user's preferred scale, common scale, chosen ratings, the scales used to give chosen ratings), within subjects design.

Subjects: Anonymous participants (187) were asked to rate their used products with all 6 rating scales, select their most preferred and most familiar scales and finally, asked to select the best-suited numeric rating score (true ratings) for each product they already rated using all scales.

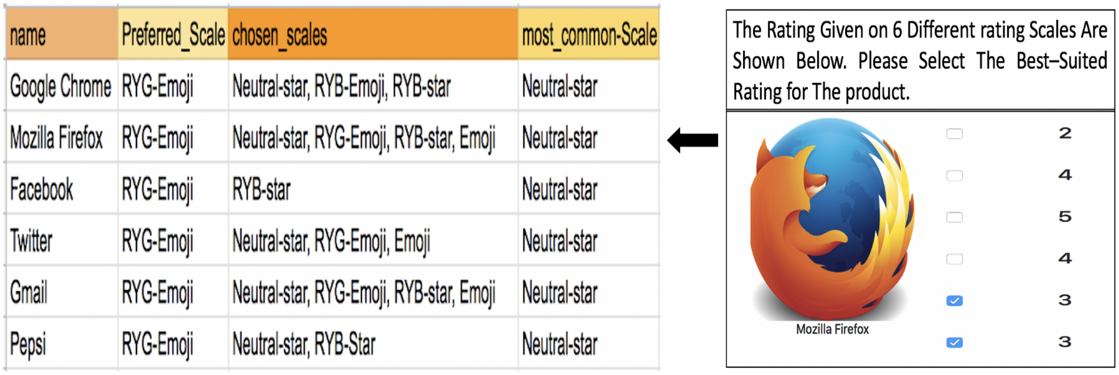
Apparatus: The raw data was collected through the user interface web-application and stored on google spreadsheets of the researches account via Google APIs.

6.2 Data Processing and Data Mining Procedure for Empirical Evaluation

Frequent data mining process called Apriori algorithm is used to mine data from users true ratings to see which rating scale is preferred for true ratings. The scales on which users gave true ratings are collected and matched with the preferred and common scale selected by same users. The process of data mining and analysis is described below.

6.2.1 Data Collection and Pre-processing: User Approved data

The raw data collected for testing the second hypothesis are the preferred scale chosen by the user from the six scales, the most common of all the scales that they see around the internet(common scale), the actual numeric ratings that they think are deserved by each product, which is called the true ratings. When users are asked to chose the true ratings, they are only shown the numeric score such as 3,4 etc. out of 5 that they gave on a scale, but the scale itself is not shown to them. The actual web page and the data collected from them to google spreadsheets are shown in figure fig:csv . The CSV file for one user is shown in figure fig:csv as an example. This user has rates 6 products using all six different scales. The user selected RYG-Emoji as his/her preferred scale and Neutral-star as common scale. The user was asked to select the best suited rating



name	Preferred_Scale	chosen_scales	most_common-Scale
Google Chrome	RYG-Emoji	Neutral-star, RYB-Emoji, RYB-star	Neutral-star
Mozilla Firefox	RYG-Emoji	Neutral-star, RYG-Emoji, RYB-star, Emoji	Neutral-star
Facebook	RYG-Emoji	RYB-star	Neutral-star
Twitter	RYG-Emoji	Neutral-star, RYG-Emoji, Emoji	Neutral-star
Gmail	RYG-Emoji	Neutral-star, RYG-Emoji, RYB-star, Emoji	Neutral-star
Pepsi	RYG-Emoji	Neutral-star, RYB-Star	Neutral-star

The Rating Given on 6 Different rating Scales Are Shown Below. Please Select The Best–Suited Rating for The product.

□ 2
□ 4
□ 5
□ 4
☑ 3
☑ 3

Figure 6.1: CSV file for a user

for each product from the ratings he/she gave. For Mozilla Firefox, the user gave 2, 4, 5, 4, 3 and 3 as ratings out of 5 using Neutral-star, RYG-star, RYG-Emoji, RYB-Emoji, Emoji and RYB-star scales respectively. When user was asked to select which of these 6 rating value is best-suited for Mozilla Firefox, user selected 3 out of 5 as best-suited /true score. The user gave 3 out of 5 on Neutral-star, RYG-Emoji, Emoji and RYB-star scales. Therefore, these three scales are stored as the "chosen_scales" for Google chrome. For the rest of the 5 products, the scales on which the true ratings are provided by the users are stored similarly. The "chosen_scales" is the ratings scales on which the users gave their true ratings. Since the "true" ratings of users used as a baseline for my hypothesis is approved by users' themselves, this approval is termed **"user-approved"** rating analysis.

6.2.2 Frequent Pattern Mining from Chosen Scales: Apriori Algorithm

Data mining is the process of discovering meaningful and interesting patterns from large amount of information [87]. A pattern is considered interesting or useful if it is proved to be valid on a test data to a certain degree and understood by humans. For testing my hypothesis, I have used the frequent pattern mining technique which is one of the widely used data mining techniques for discovering meaningful patterns appearing frequently in data.

Frequent Pattern Mining

Frequent pattern data mining is a process that searches for repeatedly occurring (i.e. frequent) item sets in a larger dataset. It leads to the discovery of interesting correlation and association rules among huge amounts of transaction records [37]. For example, if in the transaction data of customers of a superstore shows milk and bread are often bought together, than this is called a frequent pattern in the customer transaction data. The frequent pattern mining can lead to major contribution in cross-marketing, catalog design and customer shopping behaviour analysis.

Association rule: Market basket analysis for User Data

A typical and well known example of frequent pattern mining is **market basket analysis**. This process is used to analyze customer's shopping habits by finding out which items are frequently put in the "shopping basket" of customers. This analysis can help vendors make strategic shelf organization, marketing strategies etc. For instance, if the customers who are buying milk are also buying bread on the same trip to the superstore, then the this pattern can help retailers increase sales by selective marketing and organization. The bread and milk can be placed in close proximity to each other, there can be special offers arranged for both items, such as buy one milk carton to get 50% off breads. In this case, the bread and milk are frequent item set since they appear repeatedly together. To further find which of the frequent item sets are interesting, the frequent mining analysis generates association rules for the item sets. The frequent item sets which have strong association rules are considered to be interesting patterns.

For identifying strong association rules, two measures are usually considered : **Support** and **Confidence**. For my analysis, if we consider all the rating scales in user's "chosen_ scales" for giving true ratings (figure fig:csv) as a set, then the presence or absence of each rating scale can be represented by a Boolean variable(either a scale is present in chosen_scales or not). In this case, the basket is the set of chosen_scales by each user for one product. Each user have several baskets, representing set of chosen_scales for each product the user have rated. All the baskets (sets of chosen_scales for each product) for a user can then be represented by a Boolean vector of values assigned to each rating scale present in basket . So, the rating scales which appear most frequently for each user in the set of chosen_scales is derived by analysing the Boolean vectors. For example, in figure fig:csv, the user has rated 6 products, therefore there are 6 baskets or sets of chosen_scales to be analysed to find which rating scales appear most frequently in all these baskets or the frequent pattern in rating scales. These frequent patterns are represented in the form of association rules. For example, if a user having RYG-Emoji also has RYB-star rating scale as the most frequently appearing scales in their chosen_scales set, this pattern is represented by the following association rule in terms of support and confidence:

$$RYG - Emoji \Rightarrow RYB - star[support = 20\%, confidence = 60\%] \tag{6.1}$$

In the equation 4.1, support 20% for this rule means both RYG-Emoji and RYB-star scales occupy 20% of the set of chosen_scales for all products of this particular user. On the other hand, a confidence of 60% means 60% of the products (for this user) having RYG-Emoji also has RYB-star in the chosen_scales. Typically association rule is considered strong and useful if it satisfies a minimum support and a minimum confidence threshold. To find interesting patterns among the strong rules, additional correlation analysis using the parameter **Lift** can also be performed. For retrieving the most frequently appearing rating scale designs in the chosen_scales of a user, I have applied **Apriori**, an algorithm for the frequent data mining from the chosen_scales column of each participants, as shown in figure fig:csv.

6.2.3 Discovering Most Frequent Rating Scales by Using Apriori

Apriori algorithm finds frequent item sets from a given dataset of items and generate association rules from the frequent item sets. It was first proposed by Agarwal and Srikant in 1994 [6] as a faster technique for association rule generation. Apriori is used by many companies like Amazon in the Recommender System and by Google for the auto-complete features. Apriori algorithm employs a level-wise iterative search to find frequent item sets by scanning the whole database.The Apriori algorithm finds frequent item sets from a given dataset by comparing an item's frequency (support) with a threshold called minimum support i.e., check if support ≥ minimum_support [37]. After finding frequent pattern using support analysis, Apriori finds strong pattern from the selected frequent patterns i.e., items with frequent patterns which are strongly associated to each other by satisfying certain association rules. In order to do so, Apriori uses conditional probability compared with a confidence metric e.g. it checks if probability ≥ minimum_confidence to generate strong association rules.

Therefore, the basic framework for association rule generation is:

- Finding all the frequent patterns (where support ≥ minimum_support) in the dataset D.
- Generating strong association rules from the frequent patterns.

Once the frequent item sets are found, the strong association rules are generated by satisfying minimum *confidence* parameter. In order to find if these strong patterns are interesting as well, a correlation analysis is also implemented. For my analysis, I have also employed minimum *Lift* parameter as well, which finds correlation between frequent and strongly associated objects.

Apriori Algorithm

Figure fig:fcapriori shows a flow chart for Apriori algorithm and its' procedures. The input for the algorithm is the dataset D ad the minimum support value or min_sup. The algorithm generated frequent candidate item sets using the *Apriori property* . The Apriori property states that *All nonempty subsets of a frequent item set must also be frequent.* Apriori property is employed in the algorithm by two steps, join and prune.

1. **Join:** This step generates (K+1) candidate item set from K-item sets by joining each item with itself.
2. **Prune:** This step scans the count of occurrence , i.e. support count of each item in the dataset. If the candidate item does not meet minimum support or min_sup threshold,(shown in equation 4.2), then it is regarded as infrequent candidate and removed. This step is repeated until the set of K-1 item sets is null.

The algorithm scans the dataset and counts number occurrences of each item in the dataset i.e. the support count for each item. Then , the algorithm generates K candidate item sets and which satisfies the min_sup threshold by joining each candidate item sets generated during each iteration.

$$Support \quad A \Rightarrow B = P(A \cup B) \tag{6.2}$$

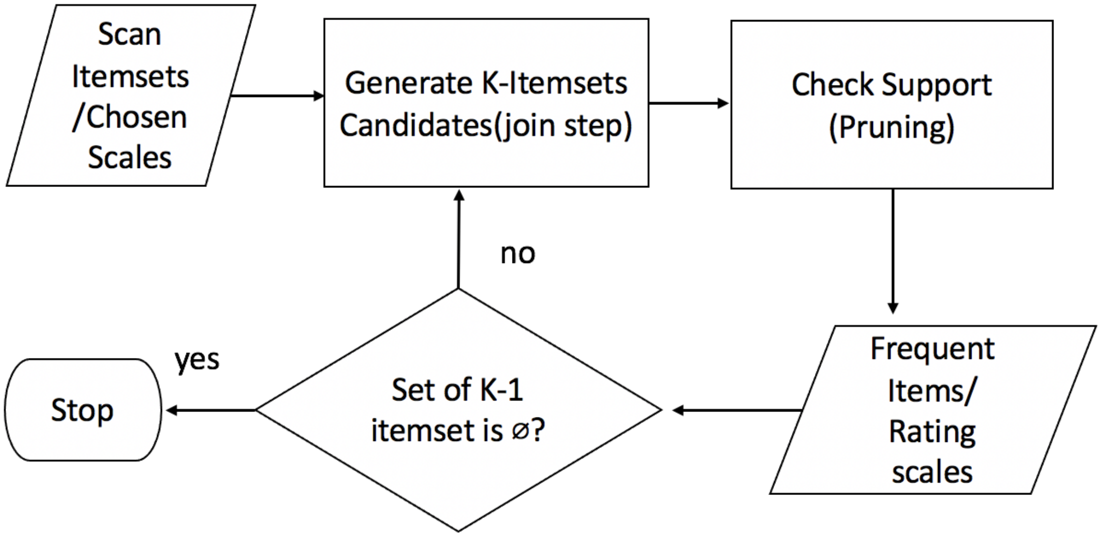

Scan Itemsets /Chosen Scales
Generate K-Itemsets Candidates(join step)
Check Support (Pruning)
no
Frequent Items/ Rating scales
Set of K-1 itemset is ∅?
yes
Stop

Figure 6.2: Flow chart of Apriroi algorithm.

Once the frequent item sets are generated, then the next step is to generate strong association rule among the frequent items. The item sets already satisfy mimimum support and now they have to satisfy minimum confidence threshold in equation 4.3.

$$confidence \quad A \Rightarrow B = P(B|A) = \frac{supportcount(A \cup B)}{supportcount(A)} \tag{6.3}$$

In the equation (1) and (2), A and B are item sets in a dataset D. The conditional probability is implemented by item set support count, where $supportcount(A \cup B)$is the number of transactions having both the item sets A and B, and $supportcount(A)$ is the number of transactions with the item set A [37]. Association rule mining usually generates a large number of rules, but a most of the time, they turn out to be redundant or they do not reflect interesting relationship among item sets. Just because one item is appearing with another item, does not always mean they are related or have any true dynamic between them. Correlation analysis can be used here to see which strong association rules are actually interesting. Therefore, for finding interesting pattern within the strong pattern sets, the frequent mining process uses correlation analysis. I have utilized *Lift* analysis to find the interesting patterns of our user reviewed dataset.

$$lift \quad A \Rightarrow B = \frac{P(A \cup B)}{P(A)\ P(B)} \tag{6.4}$$

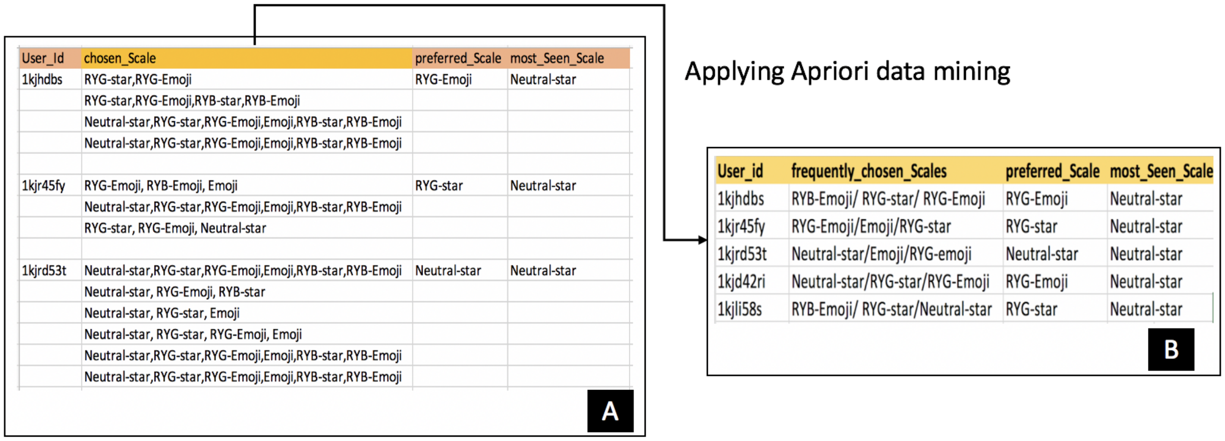
In sum, the item sets which satisfy minimum support, confidence and lift thresholds are considered to be the frequent item sets in my data.

6.2.4 Outputs from Apriori Algorithm: Frequent Rating Scale Modeling

For the second hypothesis analysis, I have applied Apriori on "chosen_scales" of each user's (figure fig:csv) data to find a set of the 3 most frequently appearing rating scale designs in the chosen_scales set for each

Table 6.1: The association rule set for few users using Apriori Algorithm

Frequently Chosen Scales	Support	Confidence	Lift
(RYG-star, RYB-star, RYG-Emoji)	0.86667	1.0	1.0
(RYG-Emoji, RYB-Emoji, Emoji)	0.97766	0.9	1.0
(RYG-star, Neutral-star, RYG-Emoji)	0.9	0.8	1.0
(RYG-Emoji, RYB-star, Neutral-star)	1.0	0.8	1.0



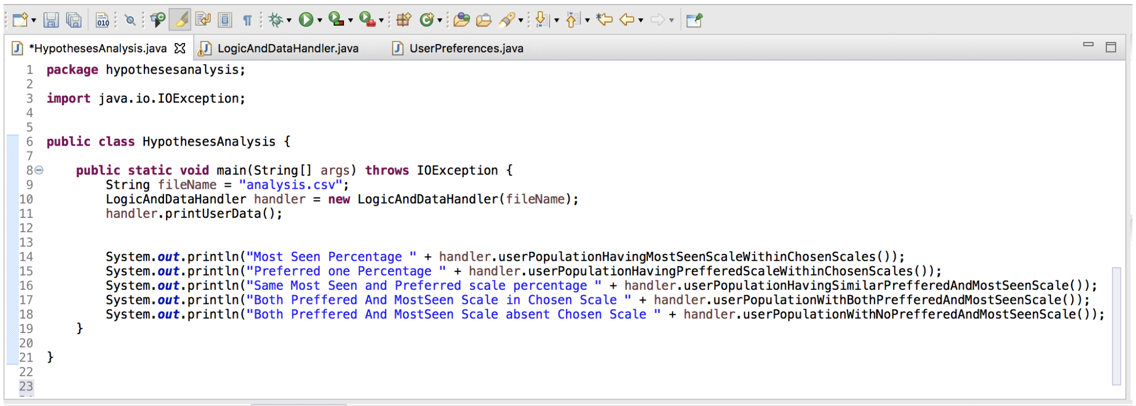
User_Id	chosen_Scale	preferred_Scale	most_Seen_Scale
1kjhdbs	RYG-star,RYG-Emoji	RYG-Emoji	Neutral-star
	RYG-star,RYG-Emoji,RYB-star,RYB-Emoji		
	Neutral-star,RYG-star,RYG-Emoji,Emoji,RYB-star,RYB-Emoji		
	Neutral-star,RYG-star,RYG-Emoji,Emoji,RYB-star,RYB-Emoji		
1kjr45fy	RYG-Emoji, RYB-Emoji, Emoji	RYG-star	Neutral-star
	Neutral-star,RYG-star,RYG-Emoji,Emoji,RYB-star,RYB-Emoji		
	RYG-star, RYG-Emoji, Neutral-star		
1kjrd53t	Neutral-star,RYG-star,RYG-Emoji,Emoji,RYB-star,RYB-Emoji	Neutral-star	Neutral-star
	Neutral-star, RYG-Emoji, RYB-star		
	Neutral-star, RYG-star, Emoji		
	Neutral-star, RYG-star, RYG-Emoji, Emoji		
	Neutral-star,RYG-star,RYG-Emoji,Emoji,RYB-star,RYB-Emoji		
	Neutral-star,RYG-star,RYG-Emoji,Emoji,RYB-star,RYB-Emoji		

User_id	frequently_chosen_Scales	preferred_Scale	most_Seen_Scale
1kjhdbs	RYB-Emoji/ RYG-star/ RYG-Emoji	RYG-Emoji	Neutral-star
1kjr45fy	RYG-Emoji/Emoji/RYG-star	RYG-star	Neutral-star
1kjrd53t	Neutral-star/Emoji/RYG-emoji	Neutral-star	Neutral-star
1kjd42ri	Neutral-star/RYG-star/RYG-Emoji	RYG-Emoji	Neutral-star
1kjli58s	RYB-Emoji/ RYG-star/Neutral-star	RYG-star	Neutral-star

Figure 6.3: Analysis of Apriroi algorithm on (A) User's Chosen_Scales scales and (B) Results of Apriori analysis

user. For most of the dataset, the selected minimum thresholds were: support ≥ 0.7, confidence ≥ 0.8 and lift ≥ 1.00. The parameters were chosen based on the user data, since I targeted to select only 3 most frequently chosen scales for each user, the threshold values were chosen to fit the target item set generation in Apriori. Table 4.1 shows a small portion of the data set (4 out of 187) processed by apriori algorithm. Apriori algorithm derived the most frequent patterns of rating scales for chosen_scales of each user. The column "Frequently Chosen Scales" presents the rating scales which were most frequently used by individual users to give true ratings for all products, according to Apriori algorithm. In the first row of the table 1, RYG-star, RYB-Star and RYG-Emoji were the most frequently used rating scale for a user to give the true ratings and the respective support, confidence and lift values (which all satisfy their minimum thresholds)for this set of frequent scales are also shown. For each user,the "Frequently Chosen Scales" can have maximum 3 rating scale deigns which satisfies the minimum thresholds decided by the apriori algorithm. The goal of using frequent pattern mining on each of the 187 users' data was to find the most frequently appearing rating scales of users used to give the true ratings on each products. In order to ensure that the algorithm delivered the strong and interesting patterns from the dataset, I used both association and correlation analysis for my hypothesis validation. The frequent data mining process is illustrated in figure fig:ap$_o utput$.

As it seen from figure fig:ap$_o utput, the “chosen\_Scales” isthesetofallratingscalesselectedbyeachuser “themselves” onwhich$



```java
package hypothesesanalysis;

import java.io.IOException;


public class HypothesesAnalysis {

    public static void main(String[] args) throws IOException {
        String fileName = "analysis.csv";
        LogicAndDataHandler handler = new LogicAndDataHandler(fileName);
        handler.printUserData();


        System.out.println("Most Seen Percentage " + handler.userPopulationHavingMostSeenScaleWithinChosenScales());
        System.out.println("Preferred one Percentage " + handler.userPopulationHavingPrefferedScaleWithinChosenScales());
        System.out.println("Same Most Seen and Preferred scale percentage " + handler.userPopulationHavingSimilarPrefferedAndMostSeenScale());
        System.out.println("Both Preffered And MostSeen Scale in Chosen Scale " + handler.userPopulationWithBothPrefferedAndMostSeenScale());
        System.out.println("Both Preffered And MostSeen Scale absent Chosen Scale " + handler.userPopulationWithNoPrefferedAndMostSeenScale());
    }

}


```

Figure 6.4: A screenshot of the Eclipse code for percentage calculation on user's preferred Scale and common Scale

6.3 Effect of User's Preferred Scale and Common Scale on User's Frequently Chosen Scales

The users have selected their preferred scales and most common scales which are the next factors for validating the second hypothesis. In figure fig:ap$_o utput, thecolumn“most\_seen\_Scale”containsthecommonscaleschosenbyusers.Apriori$

The program is developed using Java and implemented on Eclipse IDE. It reads the data file and writes separate strings for preferred scale, common scale and array of strings for the set of frequently chosen scales for each user in the data. Then it calls a Boolean function to retrieve the preferred scale present in the frequently chosen scales values of each user. The function returns true if preferred scale value is present in the frequently chosen scales array. Then it calls a separate function to count the instances where preferred scale is present in the frequently chosen scales array and returns the percentage of such instances in the total count of data in frequently chosen scales array. This same process is repeated for common scale values as well to find in how many cases both preferred and common scale are present in frequently chosen scales array. From the output of the analysis, the actual effect of user's preferred scale and common scale on user's frequently chosen scales for giving true ratings can be visualized. The results and their implications of the analysis is discussed in the next subsection.

6.4 Empirical Evidence for Second Hypothesis

The second hypothesis analyses the effect of user's preferred rating scale design on the true ratings users provide for an item, for a user-centric, customized approach for online rating model. If users tend to give the rating score that reflect how they actually feel about a product on a rating scale design that they like/prefer, this can be used as a basis for customer rating analysis model in current online marketplace. This section discusses the empirical evaluation of the first hypothesis.

Table 6.2: The results showing users giving their true ratings in preferred vs most common rating scales

The percentage of users using preferred scale for giving their true rating	**78.64%**
The percentage of users using most common scale for giving their true rating	**56.33%**
The percentage of users using both scales for giving their true rating	**34.22%**
The percentage of users using neither scales for giving their true rating	**8.6%**

6.4.1 Hypothesis 2 Validation

The user survey interface asks users to select the numeric ratings scores that they think are best fitted for the products they have already rated. The user data is prepared by selecting the scales on which the said true ratings were provided by each user as shown in figure fig:csv. These scales are termed as "chosen_scales" of each user. The data prepared for each user is analysed by using Apriori algorithm to derive the most frequently appearing scale designs in each user's chosen_scales. This is called the set of "frequently chosen scales" and it is identified as the set scales which generates true ratings from a user the most. A statistical data handler tool is used to find the percentage of preferred scale and common scale in the frequently chosen scales of each user. Therefore, this analysis reveals whether the most frequently chosen scales of users for giving the true ratings (according to their opinion) are the preferred scales or the scale they have commonly been using. The results in table 4.2 shows that major portion of the users, 78.64% have the preferred scale present in their frequently chosen scales. The commonly used scales appear in the frequently chosen scales of 56.33% users. However, there were also 34.22% users whose frequently chosen scales had both preferred and common scales. Also, a small portion of 8.6% users had neither present in the frequently chosen scales.

The statistical analysis and empirical evidence depict the viability of my proposed user preference based rating scale model in figure fig:frame. The largest portion of users have used the rating scale that they preferred to give true ratings to each products. However, the second largest population of users used the commonly seen rating scales designs to give the true ratings as well. The analysis also shows that a considerable percentage of users have used both scales to give true ratings, but this is also due to the fact that a number of users selected same rating scale as their preferred scale and common scale. The result from the analysis shows the performance of the preferred rating scale in depicting the user's "true" or "best-fitted" rating scores for each product is significant. The novel approach in my work lies in my attempt to ask users about which rating score fits well for each product and thus including "user-approved" data in the analysis. Rather than just predicting or hypothesizing what the unbiased or true rating score form the user is, I included user's own choice for rating scores, i.e. the **user-approved** true rating of each product.

The results validate the second hypothesis, since the users definitely have used their preferred scale to give ratings they themselves consider to be true for each product. The first hypothesis already established that users prefer scales that have strong visual cues. Therefore, it can be said that users give their true ratings in scales which are visually richer. This evaluation has paved the way for the next and final hypothesis of my research: to see what role the visual elements play in actual rating scores. In this case, I worked with actual rating scores users gave or **“user-given”** rating. Based on the first two hypotheses alone, the claim that UX designers should include user’s choice/preference in rating scale designs, can be supported. It can contribute to obtain the best ratings provided by users and benefit the e-commerce vendors such as Amazon or eBay to provide a credible and accurate depiction of consumer opinion regarding their products and services.

7 EFFECT OF VISUAL CUES ON RATING SCORES

The previous chapters presented the first and second hypotheses, which supported the claim that users select the rating scales which are visually informative as their preferred scale designs and also give their self-selected “true” ratings using their preferred scales. This chapter presents the evaluation of the third hypothesis by analysing whether users give positive and consistent ratings when using more visually informative scales. While previous hypotheses both focused on what role does user’s preference of rating scale play in getting “true ” ratings, the third hypothesis focuses on what role does visual cues/elements play in giving “true” ratings for each products. It considers the actual rating scores and chosen scores given by users on all rating scales for each product. Their preference or familiarity with scales and user’s individual choices are not considered in this analysis. Instead, I analysed the overall rating score of each scale for each products and the rating score selected as “true” rating by users for each product. This means the actual rating scores and how they are influenced by visual cues present in them are the basis of this hypothesis. The previous two hypotheses focused on how users are using the scale, whereas in third hypothesis, I focused on what kind of ratings each rating scale produced for each products and how the presence and absence of visual cues can influence ratings. This is termed as “user-given” ratings hereafter.

The users have often relied on visual cues provided by rating scales by assigning meanings to them [91]. These visual cues may include numeric labels to ends of scale [51], spacing between scale points [82] or even shape of the scale [26]. Certain visual cues can intensify the role of each point in the scale which can assist users as well to identify between a bad ratings and a good rating. Prior studies have successfully implied that presence or absence of visual cues have a direct impact on the rating scores of users [89], [90], [35], often causing a visible shift in the rating score. The effect of adding different labels, numbers and colours on rating scales have shown positive impact on rating scores in the work by Toepoel and Dilman [88]. My third hypothesis aims to find how the ratings are influenced by the presence and absence of rating scales and how it can be an indicator of how users perceive the cues used in my user study.

7.1 Statistical Description of Rating Scores on Six Rating Scales

Data collected in the first and third phase of user study have been used to carry out the empirical evaluation of the following findings:

Hypothesis 3: Using rating scales with more visual cues creates a positive shift and consistency in the ratings and true ratings are mostly given by users on these scales.

Design: Seven factors (6 ratings scale designs and chosen rating scores), between subject (products) design.
Subjects: Anonymous participants (187) were asked to rate their used products with all 6 rating scales and later asked to select the rating score they think is best suited for each product they already rated.
Apparatus: The raw data was collected through the user interface web-application and stored on google spreadsheets of the researches account via Google APIs.

Similar to the first hypothesis, I have also used both descriptive (Mean, Median, Inter-quartile Range) and inferential analysis (Wilcoxon test) to analyse the trend of the ratings core provided by each users on all of the 6 rating scales.

7.2 Mode, Median and Inter Quartile Range

Since the data in the user study is generated from using 5-point Likert scale, the data is ordinal. To understand what the general response trend of ordinal data, such as central tendency and spread/dispersion, median and Inter Quartile Range or IQR are used instead of mean and standard deviation respectively. Median is used to depict how positive or negative the rating scores are on each scales for each product, IQR is used for finding the consistency or ambivalence of rating scores on each scale and mode is used to analyse which rating scores are most frequently selected as the "true" ratings for each product.

The central tendency of data reveals what average respondents think or what response is likeliest to be derived from the data [49]. The spread or dispersion of the data shows whether the responses are clustered together or scattered across the range of possible values, i.e. it shows how close or far the responses are scattered across the population [49]. Usually, for continuous data with normal distribution, mean and standard deviation are used as a measure for central tendency and spread of the data. However, using mean and standard deviation for ordinal data as in this case, could produce misleading results [12]. Since my user data is ordinal, median and IQR are considered better fits for statistically describing the datasets. I have also used mode, which is the measure for the most frequently appearing data on the sample, to analyse the most popular rating score on each scale which can reveal the performance of each scale in terms of predicting the rating choices of users.

7.2.1 Median for each Rating Scale

Median is used as a statistical measure for the central tendency of the data, i.e. it shows what the 'average' participant thinks, or their 'likeliest' response. Figures fig:med1, fig:med2 , fig:med3 show the median of ratings given by all the users on each of the 21 products, using each of the 6 rating scales. The medians are shown by taking 7 products at a time due to spacing convenience. In figure fig:med1, the median of all the ratings given for KFC using Neutral-star scale is 3, the median for KFC using RYG-star is 3, the median for KFC using RYG-Emoji is 4 and so on. As evident from the median analysis, it is clear that the

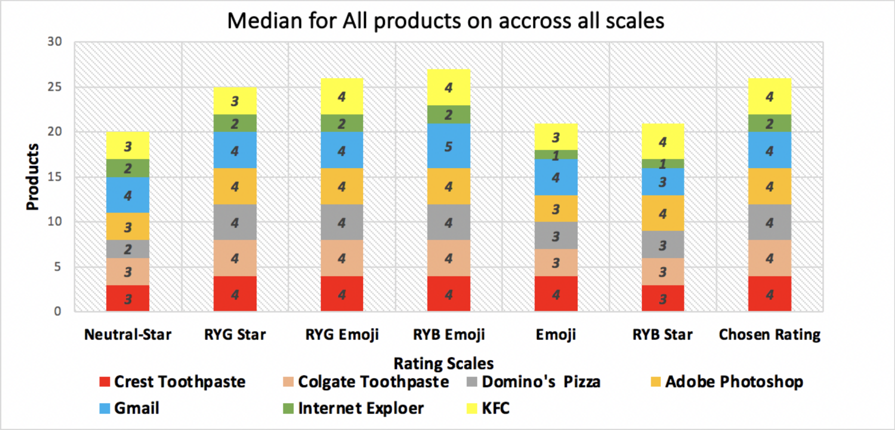


Figure 7.1: Median of ratings given for first 7 products on all 6 rating scales

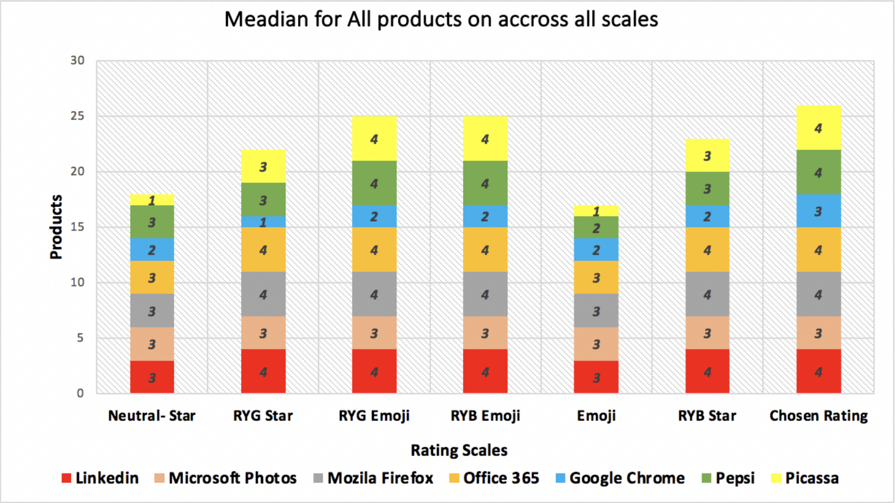


Figure 7.2: Median of ratings given for next 7 products on all 6 rating scales

average/likeliest rating scores are higher on the rating scales which has more visual cues such as RYG-Emoji and RYB-Emoji. Along with these two, the RYG-star figure fig:med2 and RYB-star (figure fig:med3) rating scales also have comparatively high rating scores, but that can vary from product to product. On the other hand, rating scale with no or minimal visual cues as Neutral -Star and Emoji have consistently lower median for all products compared to visually rich scales.

The median of chosen ratings is reflects the central trend numeric rating for a rating scale, i.e. the middle point of the ratings score distribution of each rating scale. It is evident from figures fig:med1, fig:med2 and fig:med3 that the users' chosen ratings have a higher median which is more similar to the median observed on visually informative scale such as RYG-Emoji , RYB-Emoji and RYG-star.

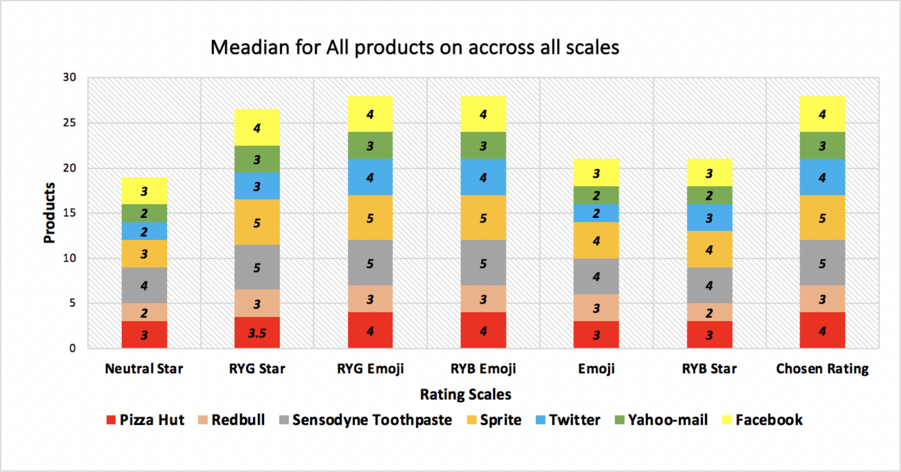


Figure 7.3: Median of ratings given for next 7 products on all 6 rating scales

7.2.2 IQR for Each Rating Scale

The IQR is a measure of spread/dispersion of the responses i.e. how strongly respondents agree with each other on a particular topic. A high IQR of a rating scale in this study, therefore, means the rating scores on that rating scale is less consistent across all the population (participants of the study), so the users are divided into different ratings for the same product using that rating scale. Some users may gave it a high score and some may gave it a low score. Subsequently, a low IQR of a rating scale means the ratings given on a rating scale is clustered together, so the users have mostly agreed upon the ratings they gave on the product using that rating scale. The users have mostly given similar ratings using this scale, which resulted in low spread/IQR score. Figures fig:iqr1, fig:iqr2 and fig:iqr3 show the IQR of all the products using all 6 rating scales. The IQR of the visually informative scale RYG-Emoji has the lowest IQR fo all products. Also RYB-Emoji and RYG-star has low IQR for most of the products except for figure fig:iqr1. The least visually rich scale Neutral-Star has the highest IQR and Emoji and RYB-star also has pretty high IQR compared to other scales. This analysis clearly depicts that users have given consistent ratings while using rating scales which have more visual cues such as RYG-Emoji, RYB-Emoji and RYG-star. The rating scales with no to minimal visual ques such as Neutral-star and Emoji has high IQR across all products and RYG-star also has pretty high IQR across almost all products.

7.2.3 Mode of Rating Scales: Frequency Analysis from User-Given Data

Mode is a statistical measure for he value that appears most frequently in a data set. observed value in a set of data and it is a measure for central tendency of a dataset. For example in our case, the mode for a product would be the most frequently given ratings for the product using each scale. If the most frequently appearing value for a product is 4, then the graph shows the number of users who gave 4 on the each scales in figure fig:data$_{m}ode.Thesamepatternisobservedincaseofthemodeofusers'ownchosentrueratingsoneachproductinfigurefig : mode$

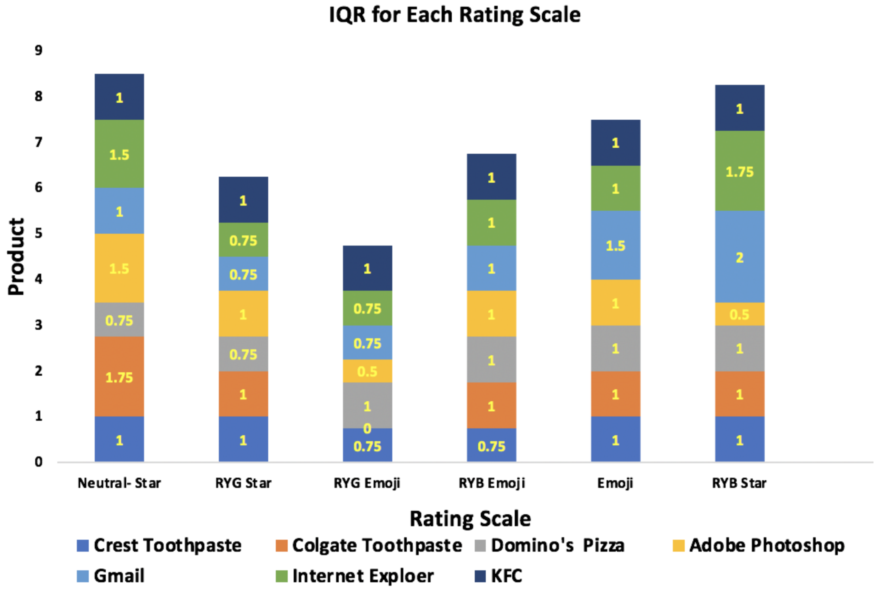


Figure 7.4: IQR of ratings given for first 7 products on all 6 rating scales

Out of all the products, I have shown the products mode which are rated most by the users. Users were first asked to select which products they have used before, the users were only asked to rate the products they selected as "used before". Figure fig:products shows the products and the percentage of users who selected them for rating. For presentation convenience, only the mode calculation process of Google Chrome, Mozilla Firefox, Gmail and Facebook are shown in this section. The overall result of mode is shown for all products. The mode for each products are represented by each 6 rating scales in two different ways:

1) Mode represented from the rating score data

2) Mode represented from the chosen ratings of users.

7.2.4 Mode Represented by Rating Scores Data

Figure fig:product$_m odeshowsthemodeforthemostpopularproductsformtheuserstudy.Themodeofgmailis4, whichmeans4out$
$starscale, 152usersonRYG-starscale, 156usersRYG-Emojiscale, 114usersonRYB-Emojiscale, 127usersonEmojiscalea$
$starscale.Thesamewayeachoftheproductsmodeispresentedinfigurefig : product_mode.Thisshowswhichoneofthe6ratingssc$

Figure fig:data$_m odeshowstheaveragenumberofusersperproductwhousedeachratingscaletogivetheratingthatisthemodev$
$star, RYG-EmojiandRYB-Emojiscalesaretheoneswhichwereutilizedmoreforgivingthemostfrequentratingsforaprodu$
$selectedratingsforeachscaleisnotafactorhere.$

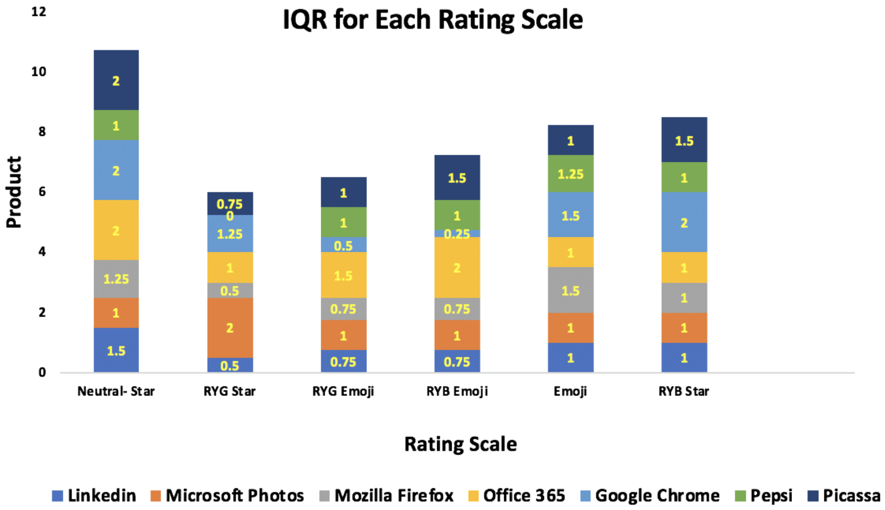


Figure 7.5: IQR of ratings given for next 7 products on all 6 rating scales

7.2.5 Mode Represented by Chosen Ratings for Each Product

After rating each products, the numeric value of each rating given on each of the six scale for each of the selected products was shown to the users and asked to select which value they think is best-suited for the product. the users selected value from their own given rating is stored and they were termed as "chosen rating" for each product. For example, if a user has rated Mozilla Firefox using 6 different scales and the ratings given on each scale were **4, 3, 4, 4, 4, 3**, then these ratings are shown to users and asked to chose the best one among them. If the user chose 4, then 4 out of 5 is the "chosen rating" of Mozilla Firefox for this user. The process is shown in figure fig:choice

The mode of "chosen rating" for each products is calculated, i.e. for a particular product, which of the 1,2,3,4,5 scores is most frequently chosen as "chosen rating"? The same process is repeated to find out which one of the rating scales can reflect the chosen rating of the product most which is illustrated in figure fig:choicemode and only the 4 products are shown here. Figure fig:choicemode shows that the mode of the chosen ratings for gmail is 4 out of 5, which means most users have chosen 4 out of 5 as the best suited rating for gmail. RYG-star scale has the highest number of 4 out of 5 rating for gmail, second highest being RYG-Emoji. From this data, it is evident that most of the time, RYG-Emoji and RYG-star has been used to give the chosen ratings of users for a product.

The total data representation for all products with all rating scales in terms of chosen ratings is shown in figure fig:modechoice. It is evident form the figure the when the mode is considered in terms of users' "chosen ratings" for each product, the RYG-Emoji and RYB-Emoji scales predicts the mode best, since these two scales have the highest average of users giving the mode value as rating scores for each products.

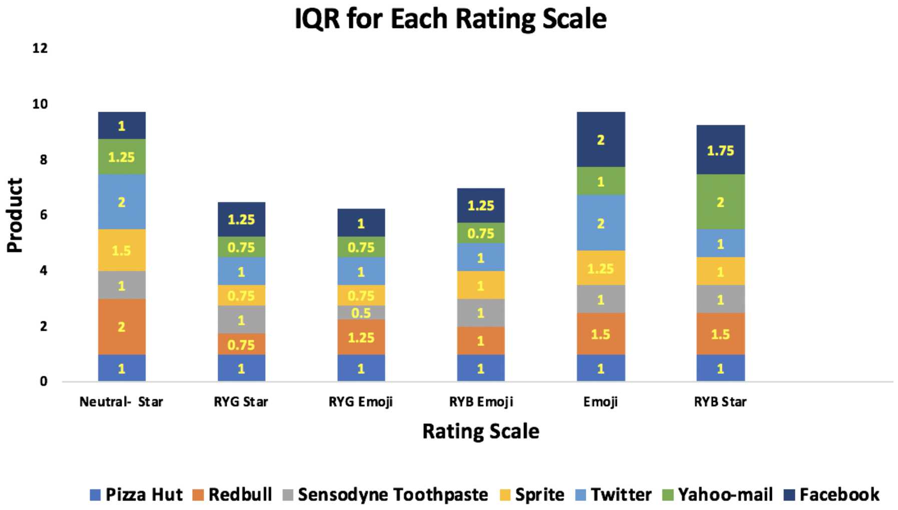


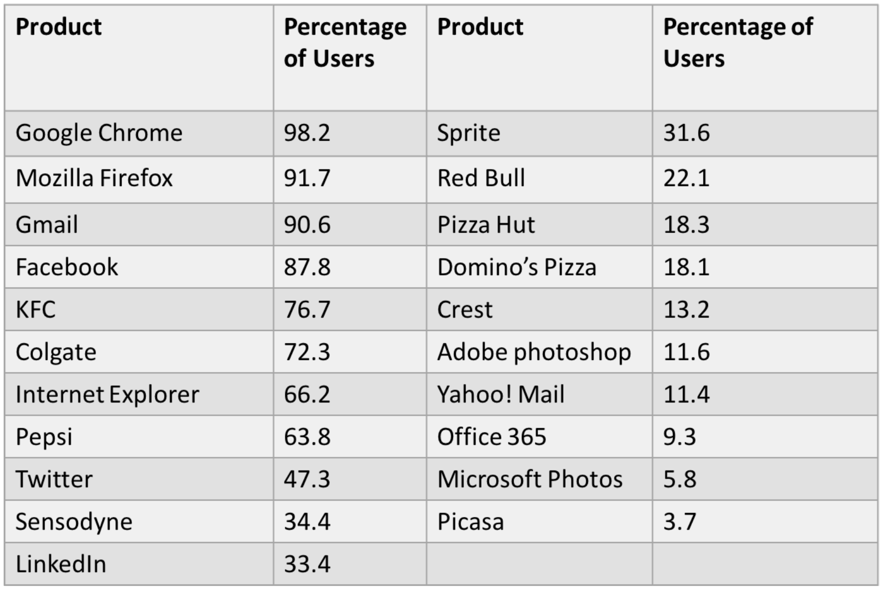
Figure 7.6: IQR of ratings given for next 7 products on all 6 rating scales

7.3 Result and Interpretation for Hypothesis 3

The result from third hypothesis validates the claim that users give comparatively positive ratings on visually informative scales. It also shows that users give consistent ratings while using visually rich scales for a product. Additionally, the analysis revealed that the ratings scores given on visually rich scales are mostly chosen as "true" ratings for each product. These findings and results results are discussed below.

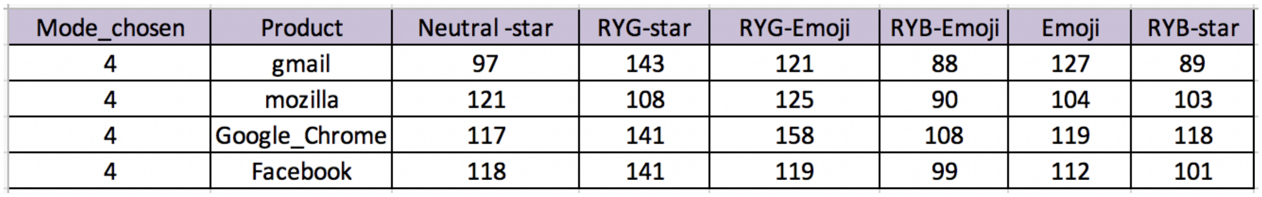
7.3.1 Major Findings in Median Analysis

From the median analysis, depicted in figures fig:med1, fig:med2 and fig:med3 the central tendency of rating scores for each scale for each product is shown. It each colour block shows the median for a product on the ratings scales. For instance, in figure fig:med1, the median of all the rating scores given by all users who rated crest toothpaste using each scale is shown by using the yellow block. The median of ratings given for Crest toothpaste is 3 on Neutral-star and RYG-star, 4 on RYG-Emoji, RYB-Emoji and RYB-star. The chosen rating values are the user selected true ratings of crest toothpaste and the median of the chosen rating for crest toothpaste is 4. If all products in all three figures are observed, it is seen that in most cases, the median of rating scores for a product is comparatively higher in visually rich rating scales, specially in RYG-Emoji, RYB-Emoji and RYG-star scales. On the contrary, the rating scales which offers less visual cues such as Neutral-star and Emoji have low median scores for almost all products. The most interesting observation that can be made here, is that the user chosen rating for each product (shown using the same colour block) always has the higher median value, just like the ones seen in RYG-Emoji or RYB-Emoji rating scales. Another finding from median study is the fact that RYB-star has also yielded lower median value despite offering visual cues using colour scheme (warm-cool).



Product	Percentage of Users	Product	Percentage of Users
Google Chrome	98.2	Sprite	31.6
Mozilla Firefox	91.7	Red Bull	22.1
Gmail	90.6	Pizza Hut	18.3
Facebook	87.8	Domino's Pizza	18.1
KFC	76.7	Crest	13.2
Colgate	72.3	Adobe photoshop	11.6
Internet Explorer	66.2	Yahoo! Mail	11.4
Pepsi	63.8	Office 365	9.3
Twitter	47.3	Microsoft Photos	5.8
Sensodyne	34.4	Picasa	3.7
LinkedIn	33.4		

Figure 7.7: Percentage of users for product selection and ratings



Mode_chosen	Product	Neutral -star	RYG-star	RYG-Emoji	RYB-Emoji	Emoji	RYB-star
4	gmail	97	143	121	88	127	89
4	mozilla	121	108	125	90	104	103
4	Google_Chrome	117	141	158	108	119	118
4	Facebook	118	141	119	99	112	101

Figure 7.8: A display of mode for most popular products based on rating score data

RYG-Emoji and RYB-Emoji scales have the highest median value, with RYG-star in the second highest median. These scales offer both visual cues : colour and icon where as RYG-star offers only colour metaphor. RYB-star had lower median which can be connected to the fact that only using warm cool metaphor may not be enough to draw users' attention since this is not a very popular colour metaphor and it is more subtle. However, using warm cool metaphor with emoji has drastically changed the score for RYB-Emoji. On the other hand, the traffic -light metaphor is widely used across the world which may have acted as a factor while giving higher rating score using only the colour scheme in RYG-star scale. The major takeaway from the median of all rating scale is that users have definitely been influenced by the visual cues offered by the rating scales which reflected on their rating scores. The rating scales with no or minimal/lesser known visual cues such as Neutral-star, Emoji, RYB-star has lower median. Therefore, using visual cues can elicit positive ratings from users, while using less visual cues can shift the users' rating scores to a more middle or lower end of the rating scale.

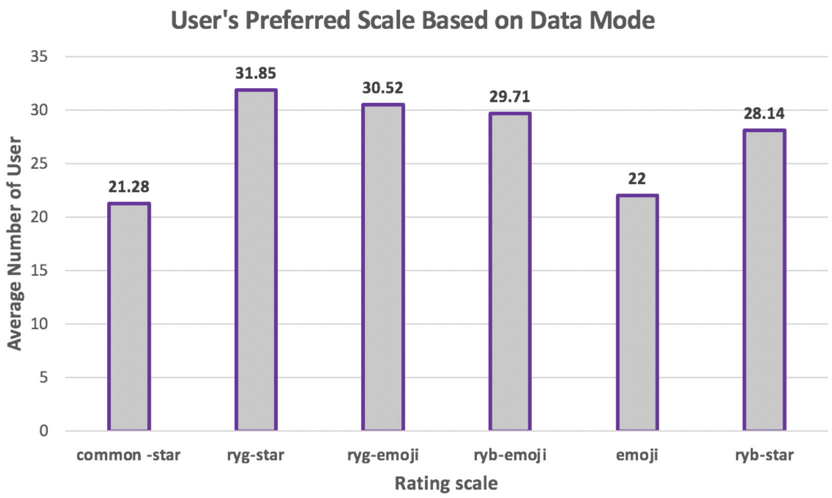
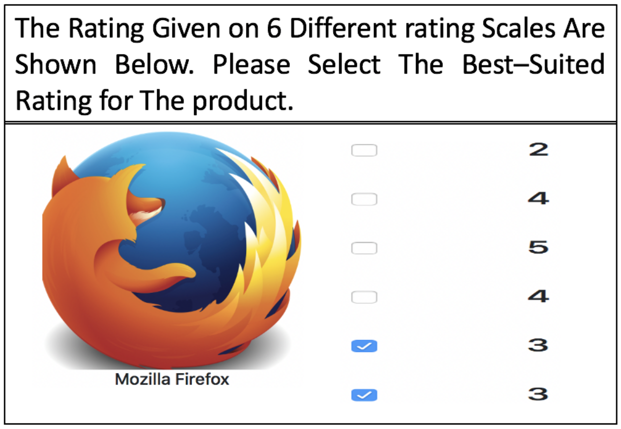


Figure 7.9: Mode for all products from user rating data



The Rating Given on 6 Different rating Scales Are Shown Below. Please Select The Best–Suited Rating for The product.

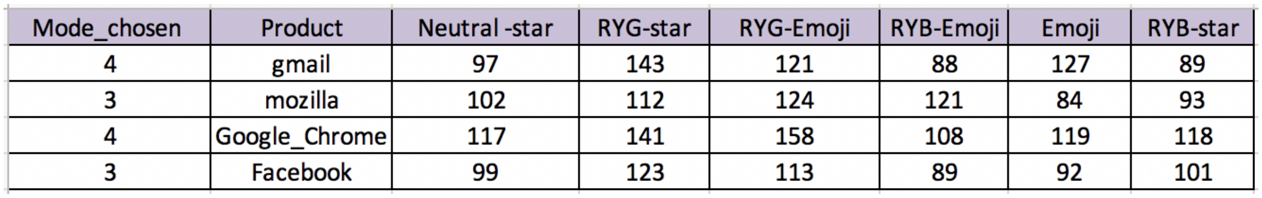
Mozilla Firefox

□ 2
□ 4
□ 5
□ 4
☑ 3
☑ 3

Figure 7.10: Users selecting the chosen rating for Mozilla Firefox

7.3.2 Major Findings in IQR analysis

The IQR analysis in figures fig:iqr1, fig:iqr2 and fig:iqr3, show the consistency of the ratings given on each scale for each product in the same manner as median, by using individual colour blocks for each products. In figure fig:iqr2, the IQR value for ratings given to LinkedIn using Neutral-star scale is 2, which means the ratings given for LinkedIN using this scale can vary with an average difference of 2 points from user to user. For RYG-star, the IQR is 0.75, for RYG-Emoji and Emoji it is 1, for RYB-Emoji and RYB-star it is 1.5. These values suggest that the difference of users' ratings given on LinkedIn while using less visually rich scales such as Neutral-star is pretty wide. However, the user ratings using RYG–Emoji, RYG-star have less IQR, which means user ratings differed less with these scales. The total IQR analysis shows that users gave same ratings or least diverse ratings using RYG-Emoji rating scale. Users' ratings varied widely across Neutral-star and Emoji scales on the same product. Ratings on RYB-Emoji and RYG-star have been less



Mode_chosen	Product	Neutral -star	RYG-star	RYG-Emoji	RYB-Emoji	Emoji	RYB-star
4	gmail	97	143	121	88	127	89
3	mozilla	102	112	124	121	84	93
4	Google_Chrome	117	141	158	108	119	118
3	Facebook	99	123	113	89	92	101

Figure 7.11: A display of mode of chosen ratings for most popular products

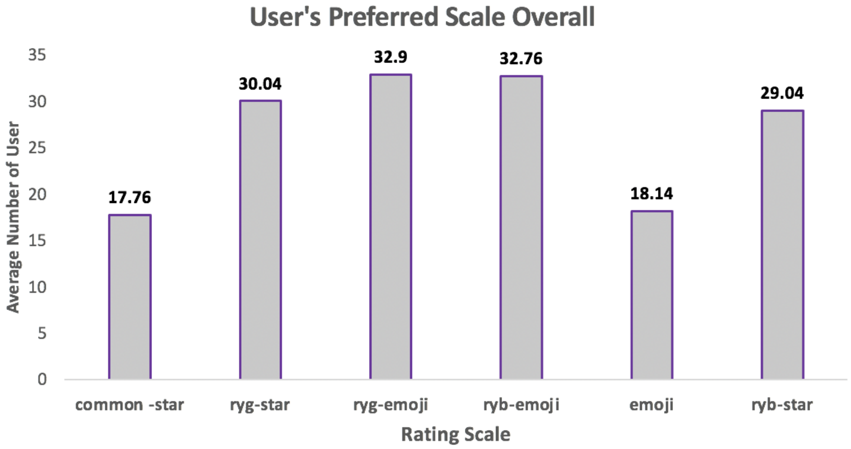


Figure 7.12: Mode of chosen ratings for all products

diverse than those on Neutral-star and emoji scale, but way less consistent than RYG-Emoji scale.

Therefore, the IQR analysis revealed that when using the visually informative scales, users tend to give consistent ratings, rather than when they are using a less visually rich scale. The visual cues offered in the rating scale RYG-Emoji or RYB-Emoji can help users make a stronger decision regarding what rating to give a product and the users tend to be consistent in their rating behaviour more. On the other hand, when using a rating scale that offers them no visual aid in rating process, users tend to give random ratings to the product using that scale, which eventually results in inconsistent ratings. For instance, a test case can be made in terms of the product Internet Explorer rating scores. The median analysis in figure fig:med1 showed that the average rating score for internet explorer is lower than other products, i.e. users do not like this browser. In the IQR analysis of Internet Explorer on figure fig:iqr1, it is clear that this low rating is pretty consistent while using RYG-Star and RYG-Emoji scale, (they both have IQR 0.75). Therefore, it can be stated that when rating scales offer more visual cues, users give strong and consistent opinions about the product they are rating using that scale.

7.3.3 Mode analysis: Data Mode and Chosen Rating Mode

Mode analysis is shown in two cases: considering users; chosen ratings (user's choice perspective) and considering only the rating scores from data (rating score data perspective). From the rating scores data perspective in figure fig:data$_{m}ode, itisevidentthatthemostcommonratingscoreforaproductisusuallyobtainedontheRYG - Emoji, RYG-StarandRYB-Emojiratingscales.Thismeansthattheseratingscaleswereemosteffectiveincapturingthemost$

Then from the user's choice perspective, the same mode analysis is conducted on the user's own chosen ratings for a product, to see which of the rating scales captured the user's self-selected "true" ratings for each product best. It is evident in figure fig:modechoice that RYG-Emoji and RYB-Emoji are the two scales where the chosen score for a product was given. Therefore these two scales were most effective in capturing the "true" opinion/review about a product according to users. The users ratings, be it the actual most common ratings score for a product and the true ratings of a product, both are captures well by the the rating scales which offered more visual cues. On the other hand, the rating scales with less visual cues such as Neutral -star and Emoji have poor performance in both cases for predicting users' popular or true opinion/ratings of a product.

Table 7.1: Wilcoxon signed rank test for ratings on all products

Pairwise Comparison	**Z-Value**	**p-Value**
Neutral-Star and Ryg-Emoji	-6.56	**0.000**
Neutral-Star and Ryb-Emoji	-4.12	**0.000**
Neutral-Star and Ryg-star	-4.38	**0.000**
Neutral-Star and Ryb-star	-2.45	0.042
Emoji and Ryg-Emoji	-3.18	**0.000**
Emoji and Ryb-Emoji	-5.67	**0.000**
Emoji and Ryg-star	-3.89	**0.000**
Emoji and Ryb-star	-2.13	0.023
Neutral-Star and Emoji	-0.78	0.322

7.3.4 Validation of Difference of Ratings on Each Scale: Wilcoxon Ranks Test

In order to validate the findings, Wilcoxon signed rank test was conducted on the ratings of each rating scales for each product. The results showed significant differences among ratings given using six different scales for most products except Internet Explorer, Office 365 and Microsoft photos. For these three products, the null hypothesis was not rejected (p >0.005) for Neutral-star and Emoji rating scales. This means for these products, the ratings given on Neutral-star and Emoji did not have any significant difference. The results of the statistical analysis of the ratings given on visually rich scales RYG-Emoji, RYB-Emoji, RYB-star and RYB-Emoji with less visually rich scales Neutral-Star and Emoji are shown in table 7.1. The highlighted values are the pairs which showed significant differences.

The results from the statistical and empirical analysis can support the third hypothesis stating that users using a visually informative scale can elicit positive ratings from users compared to less visually effective scales. The rating scales which offer visual cues to assist users can influence users to give solid and consistent ratings for the products as well as capture the most popular and true opinion of users much better that rating scales which does not provide any visual aid to users.

8 DISCUSSION AND CONTRIBUTION

The three proposed hypotheses in this research collectively form the foundation of a user-preference oriented approach for the rating collection system in recommender and e-commerce systems. The UX designers generally take a common one-size-fits-all approach when it comes to providing rating tools to users whereas these ratings can make or break a product/service in today's climate. Most UX or UI design elements base the rating system on what seems to be the popular choice, rather than customizing the rating tools based on how users want it. This arbitrary nature of rating scales pose a serious user experience (UX) problem [3]. The proposed rating system works both ways: it allows users to choose their own rating scales with visual cues and colour shading that makes them more comfortable in rating process which in turn can elicit the true /accurate ratings from them. On the other hand, the use of meaningful icons and colour metaphors creates a consistency and positive shift in the ratings which means their ratings are more reliable and less negative, which is very useful for the online product sale and recommendation process. The results and contribution of research implications of my thesis are discussed in this chapter.

8.1 Hypothesis Validation

The workflow of the hypothesis validation is shown in figure fig:wf.

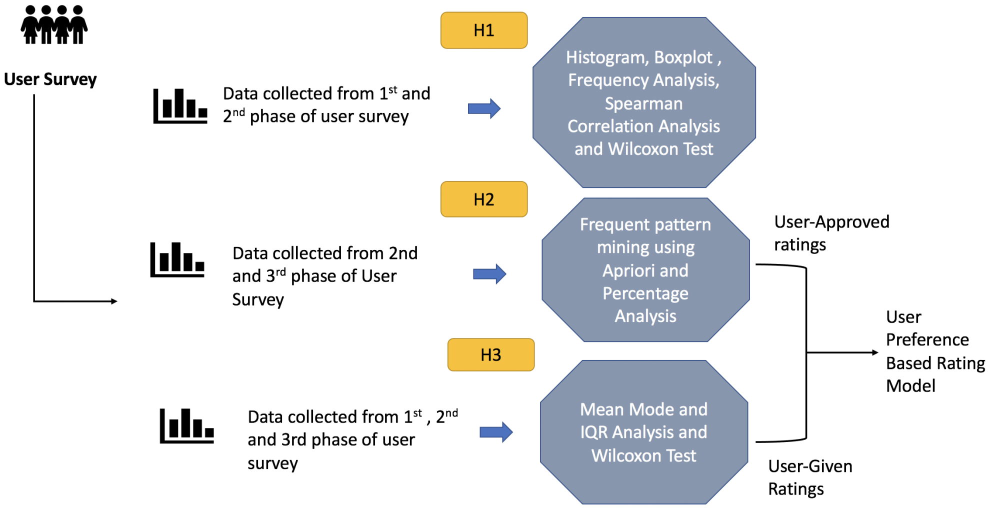

User Survey
Data collected from 1st and 2nd phase of user survey
H1
Histogram, Boxplot , Frequency Analysis, Spearman Correlation Analysis and Wilcoxon Test
Data collected from 2nd and 3rd phase of User Survey
H2
Frequent pattern mining using Apriori and Percentage Analysis
User-Approved ratings
User Preference Based Rating Model
H3
Data collected from 1st , 2nd and 3rd phase of user survey
Mean Mode and IQR Analysis and Wilcoxon Test
User-Given Ratings

Figure 8.1: Workflow of research process and inference

The three hypotheses proposed in my work have all been validated using different statistical and empirical tools. The validated hypothesis and their contributions are shown below:

8.1.1 Validating Hypothesis 1

Users prefer a rating scale which provides visual cues that assist users in their rating process, overrating scales they most familiar with on the internet.

The validation of first hypothesis was based on data analysis from the ratings of users on each scale and the rating scale selected as preferred and most common. The following steps were takes to validate the data:

1. The boxplot and histogram analysis of all the ratings collected on the preferred and most common rating scale selected by users in figure fig:box and fig:hist.

2. Spearman correlation analysis on the rating scores provided on each scales in table 5.1.

3. Finally Wilcoxon test was performed on each rating scales to validate the difference between ratings.

8.1.2 Validating Hypothesis 2

Users provide their true ratings when using rating scale designs they prefer themselves.

The validation of second hypothesis was based on frequent data mining and association rule mining from the user-approved ratings for each product, the rating scales on which chosen ratings were provided and the preferred and most common scale chosen by users. The following steps were takes to validate the data:

1. The data collected from user survey with chosen ratings and chosen rating scales were analysed using Apriori algorithm to derive the most frequently associated rating scales for each user as shown in table 4.1 and figure fig:ap$_{o}utput.EmpiricalanalyticaltoolusedonthesetofchosenscalederivedfromthefrequentpatternminingusingAprioiandthe$

8.1.3 Validating Hypothesis 3

Using rating scales with more visual cues creates a positive shift and consistency in the ratings and true ratings are mostly given by users on such scales. The validation of third hypothesis was based on data from actual rating scores of users given on each products on each rating scale. The following steps were takes to validate the data:

1. The data collected from user survey are analysed using descriptive statistics median, IQR and mode analysis, results shown in figure fig:med1, fig:med2, fig:med3, fig:iqr1, fig:iqr2, fig:iqr3, fig:modechoice and fig:data$_{m}ode.FinallyWilcoxontestwasperformedonratingsprovidedoneachscaleforeachproducttovalidatethedifferenceof$

2. **Research Question 1:** *Do users prefer the rating scale they are most accustomed to, or do they prefer scales which have more visual cues?*

   The users were asked to select which of the 6 scales would they like most and which one they have seen/used most on internet. The results show that although 86.4% of users commonly saw the Neutral-star rating scale online, 54.75% preferred RYG-Emoji scale as the preferred one, as opposed to 21.08% who choose the Neutral-star as their preferred scale. The rating scale RYG-Emoji has both visual icon and colour metaphor which serve as visual cues to ignite emotional response in users to help in their rating process.

Research Question 2: *Do users rate differently on the scale that they prefer than on the scale they are most accustomed to? If so, how significant is the difference of rating scores they give using the two scales?*
The histogram and box plot clearly show that the ratings given on the preferred scale of users and most commonly selected scales are different on same products. The correlation analysis between two ratings on preferred scale and most common scale shows a weak correlation (0.249519) which indicates that the ratings on these scales are not dependent on each other.

Research Question 3: *Do users give the rating that they truly believe the product deserves (called hereafter true rating) on a scale that they prefer?*
Frequent data mining pattern analysis using Apriori algorithm and statistical evaluation tool showed that users selected the ratings (which users themselves think are best-suited for each product), are for the most parts, given by using rating scales the users preferred. The percentage of of users who gave their true ratings on their preferred scale in 78.64%, whereas the percentage of of users who gave their true ratings on their most commonly used scale is 56.33%.

Research Question 4: *Do users give their true ratings for a product using scales with more visual cues than scales which have less visual cues?*
The research question 4 can be derived from the answers from research question 1 and 3. From question 1 , it is demonstrated that users have selected rating scales with visual cues as their “preferred” scales. From question 1, it has been deduced that users clearly gave their own approved best/“true” ratings on scales they prefer. Therefore, it can be concluded that users tend to give true/best fitted ratings when they are using ratings scales with more visual cues, such as RYG-Emoji, than neutral monotone scales such as Neutral-star.

Research Question 5: *Does using rating scales which contain more visual cues impact the consistency of ratings given by users and cause any positive /negative shift of the ratings?*
The median and IQR analysis for the ratings given on each scale for each product show that when users are using rating scales with visually rich connotations such as RYG-Emoji, RYb-Emoji scales, the ratings tend to have a slight positive shift compared to scales which offer no distinct visual cues (Neutral-star) or little visual cues (Emoji). The users are more critical of a product or they tend to stay in the neutral or middle ground when using a rating scale that offers no visual cues to help them. It can be stated that red-yellow -green or traffic-light scheme is a lot more familiar to users than red-yellow-blue or warm-cool colour scheme, which is more subtle and works on a more subdued cognitive level than traffic-light metaphor. This can be a possible reason why users tend to allow RYG-Star to have a prominent effect than RYB-star, even though they have same icons. However, when RYB-Emoji scale is use, both colour and emoji icon together have influenced users nearly as much as RYG-Emoji scale. The IQR analysis showed that when users are using visually rich scales, their ratings given to a particular product stayed consistent compared to less visually

informative scales. On the other hand when they are using less visually rich scales, their ratings are less consistent across the same product, which is an indication that using rating scales which does not reduce the cognitive load of users can cause users to give random ratings which are not reliable.

Research Question 6: *Does using rating scales which have visual cues impact the most frequent rating value given on a particular product?*
The mode analysis of each rating scale shows that the most frequent rating score for a given product is mostly given on using rating scales which offer more visual cues such as RYG-Emoji than less visually informative scales such as Neutral-star scale. For instance, if the most commonly given rating score for Facebook is 4 out of 5, then this 4 out of 5 rating is most given on RYG-Emoji scale. On the other hand, the most commonly given rating on Neutral-star scale is 3 out of 5. This can conclude that the visually rich scales can predict the popular rating choice of users better than other generic scales.

8.2 Research Contributions and Recommendations

Online ratings are used by vendors to understand user interest and demand and helps improve the quality of service. Inaccurate depiction of user feedback can lead to misinterpretation of user interest and result in inefficient services on system's part. The credibility of user rating depends on how seriously or actively users participate in rating process. The visual presentation of rating scale has significant impact on the rating behaviour of users. To this best of my knowledge, this thesis is the first research to examine the role of user preference in rating scales for generating the true or best-fitted ratings from users. The thesis also provides an in-depth analysis of how using strong visual cues can affect users in selecting the true ratings (according to users themselves) for an item. The following insights can be stated as contributions from my research that can be utilized for rating system design:

1. **Improving Current Rating System by Considering User Preference in Choosing Rating Scale:** E-commerce systems, recommender systems, online marketplaces depend on customizing user's needs based on their personalized choices. The research suggests that the online vendors should also consider including user's choice when it comes to rating scales. My analysis shows that the current rating system, where a one-size-fits-all rating scale design approach is used is not necessarily what users like and it can degrade the quality of user opinion and feedback collected by the system.

2. **Reliable Feedback Leading to Better Customer Service:** Findings from my research show that most users give ratings they truly think a product deserves when they are using a rating scale of their own choice. Therefore, in order to build a credible and useful feedback mechanism, users should be allowed to rate using the scale of their own choice, not what the system wants them to use. Giving true feedback eventually results in feedback that provides real insight into users interests and opinions. This is a scope for online vendors to improve product quality or service in accordance with what users

need. If personalized ratings system can mitigate different biases associated with user rating, it paints a clearer picture of actual user feedback. UX designers should employ mechanisms to allow users to choose the rating scale of their choice. This will improve the user feedback quality and in turn, enhance the efficiency of online businesses and user satisfaction.

3. **Visual Cues to Assist Users in Choosing Best Ratings:** The study shows when it comes to including options to give to users for rating scales, visual cues which have emotional connotations are particularly useful. For example, when users are using a scale with simple star icons, they do not have any emotional response to it, which can translate to the ratings they give as well. They tend to be less attentive to what ratings they are giving or give negative ratings without much thought. When the stars are replaced by human emojis, which has emotional elements, users can relate to it and steer away from insincere or extremely negative ratings. Similarly, when a widely popular colour metaphor such as red for negative, yellow for medium and green for positive is used on rating scale, it makes users more aware of each rating value on a scale. The research shows that users tend to choose such visually informative scales more than other options, which in turn helps them giving true ratings for a product. Even when the preference of scale is taken out of picture, the users chose the ratings on visually rich scales to be the true rating for the products. Therefore, including rating scales with meaningful visual elements can be effective for coaxing users into giving less negative and more sincere ratings.

4. **Visual Cues in Rating Scales Reduces Negative WOM:** The results from research showed that when rating scales have more meaningful visual cues such as colour metaphors or emoji, the ratings on such scales tend to be more positive and consistent across the products. Since WOM in the main factor in success or failure of online marketplace, using meaningful visual cues in rating scales can help reduce extreme negative ratings from users.

5. **Demonstrating that Online Seller and Buyer Both are Benefited from a Preference-based Rating System Approach:** Including user preference in rating system and meaningful visual elements in rating scales can serve both parties included in online business and recommender systems. Giving users a personalized rating scale mechanism can assist them in expressing the real opinion about their interests and experiences. Their reliable feedback in turn can improve the customer service, product quality and build trust between buyers and sellers. Large volume of negative ratings is one of the key factors that can affect consumer's buying decisions and hamper the trust they have on the seller. Therefore, using visual cues to reduce extreme negative ratings can help online vendors in making necessary modifications without losing consumers due to negative social influence bias.

9 CONCLUSION, LIMITATIONS AND FUTURE WORKS

The user study designed for this thesis used six different rating scale deigns to elicit ratings of users for products they have used before. The study then asked users to select their preferred scale and commonly used scales and select the numeric ratings they think suits the product best. The results and analysis demonstrated that users prefer rating scales with visual cues than neutral rating scales and this preference contributes in obtaining true ratings from them as well. The true ratings of users for a products are assessed based on users own selected or approved rating scores. This finding is significant in terms of getting reliable user feedback, which is indispensable for successful online business ventures. Including user preference in rating scales for e-commerce and online recommender systems can improve user engagement in rating process and including meaningful visual cues can ensure accuracy in terms of getting useful user feedback.

9.1 Limitations

Despite the contributions of the thesis, the work also has some limitations.

1. The study goals were to investigate the effect of individual user choices regarding rating scale designs but it did not consider the participants demographic and personality traits into account. Studies already explored the role of personality traits and cultural background on rating [46]. Age and gender can also play a role in persuading users [72]. The study suggests that users should get chance to select rating scale deign they like and rating scales should have strong visual cues to help users give the best ratings. Since users have to choose from a limited number of options of rating scale designs, it is important to include users age gender , cultures as important factors to understand what kind of visual design is better fit for rating options for users.

2. The experiment was conducted on users rating products they have already used before , which means the users might already have a strong opinion of the products performance which can be rigid despite using different rating scales. Recalling the actual ratings they have in mind might be a bit easier, which can make the actual effect of different visual cues of rating scales a bit biased. Previous literature already showed that the objects being rated [17] and the previous experience of users with ratings [61]. Using a rating task for something they not experienced before, such as a joke or an image measuring how funny or attractive it is respectively can give a better insight on the actual effect of rating scale.

3. The thesis used familiar metaphors such as traffic lights and emojis, which is not completely novel rating scale design to users. By using an altered version of existing rating formats such as using green as negative and red as positive colour, the effectiveness of visual cues could have been better analysed.

4. The effect of using visual cues in rating scales only conducts a quantitative analysis of user survey. The actual insights for preferring a certain scale and which feature of the scale (icons or colour) appealed more to the could have acted as a qualitative element to the validation of proposed framework.

These limitations can be addressed with further development and extended research on the same proposed framework and the future works related to this thesis are discussed in next section.

9.2 Future Work

The research works and analysis in this thesis provided an understanding and awareness of the relationship between user preference and true rating scores, as well as the role of visual cues in both of them. The future works added to the extension of this proposed framework can be as follows:

1. **Customizing Rating Systems based on Demographics:** The role of certain demographic features of users such as age and gender in the rating behaviour of users while using rating scales can be the next step to create a truly adaptive customized rating system. It has already been explored that age and gender play a vital role when it comes to emotional and social persuasion in [72]. Young people may prefer more colourful presentation while matured participants can prefer verbal labeling. It will be interesting to investigate how users are going to react to the rating model when tailored to their demographics as opposed to the randomization approach employed in this study.

2. **Experiments with Unforeseen Visual Cues:** It will be certainly an interesting to use novel and unforeseen visual features in rating scales such as using red as a positive hue and green as a negative hue. This can help identify the contrast between the effect of visual cues and actual feeling of users.

3. **Tailoring the Rating model on Participants Personality:** The personality of users play a strong role in the rating scores [16]. Same visual cues can represent a different aesthetic to people of different personalities. Tailoring the rating scale model study to consider user personality can add to the efficiency of system more.

4. **Follow-Up Study with Diverse Itemset:** Current study in my research only uses daily common items for user rating. Online services are available for every aspect of life such as car-sharing (Uber, Lyft), food-delivery(Uber Eats, Foodpanda), traveling (Yelp) and so on. A follow-up study with different kinds of products and services can help identify how preference of rating scale and visual cues can vary depending on the object being rated.

The findings and contributions of this thesis can lead avenues for user-centric approach in rating system. The designs and results discussed here build off of a vast amount of practical and theoretical prior literature, and I hope that my work can serve as a launch pad for UX designers and online merchants in advancing the field even further and serve purpose of building customized online feedback mechanism.

Online WOM is the key factor in product sales in e-commerce and recommnder systems, and the WOM manifests from collecting user ratings. Still, online rating systems require attention from UX designers and online service providers given this is what attracts and promotes products to users in online spheres. Persuading users to leave useful ratings that reflect their interest and experience can help both the parties involved in online marketplace. This thesis proposes a user preference based rating scale framework that can provide insight and initial guideline to UX designers for future development in rating systems. The thesis adapted a novel approach conducting experiments with user's own approved rating outcomes to identify which ratings are actually useful and true to what users feel about products. This approach supports the validity of my proposed hypotheses, both in terms of user's conscious opinion and user given ratings. Both *user-approved* and *user-given* data showed similar outcomes, as in user giving true ratings by using preferred scales and strong visual cues getting the preference of users. Rating scales using visual cues were preferred by users, and user-preferred scales elicited the true ratings of users; this finding is a step towards building more credible online commerce sites. More valid and reliable user ratings equal useful WOM, which brings out scopes of improving service, which eventually results in better customer service, thus increasing product sales. These findings can contribute as an initial guideline and offer interesting insights which can lead the way for future development in rating scale systems for e-commerce and recommender systems.

References

[1] Global trust in advertising and brand messaging. *https://www.eaca.eu/wp-content/uploads/2016/06/Global-Trust-in-Advertising.pdf*. Accessed: 2020-05-27.

[2] New data: 97% of consumers depend on reviews for purchase decisions. *https://www.powerreviews.com/events/consumers-depend-on-reviews/*. Accessed: 2020-05-27.

[3] The user experience (ux) of rating things. *https://www.bennadel.com/blog/2509-the-user-experience-ux-of-rating-things.html*. Accessed: 2020-09-30.

[4] Jill L. Adelson and D. Betsy McCoach. Measuring the mathematical attitudes of elementary students: The effects of a 4-point or 5-point likert-type scale. *Educational and Psychological Measurement*, 70(5):796–807, 2010.

[5] Gediminas Adomavicius, Jesse Bockstedt, Shawn Curley, and Jingjing Zhang. Reducing recommender systems biases: An investigation of rating display designs. pages 112–116, 02 2019.

[6] Rakesh Agarwal, Ramakrishnan Srikant, et al. Fast algorithms for mining association rules. In *Proc. of the 20th VLDB Conference*, pages 487–499, 1994.

[7] Duane F Alwin. Margins of error: A study of reliability in survey measurement. 547:934–946, 2007.

[8] Georgios Askalidis, Su Jung Kim, and Edward Malthouse. Understanding and overcoming biases in online review systems. *Decision Support Systems*, 97:97–103, 03 2017.

[9] Georgios Askalidis, Su Jung Kim, and Edward C Malthouse. Understanding and overcoming biases in online review systems. *Decision Support Systems*, 97:23–30, 2017.

[10] Christopher Avery, Paul Resnick, and Richard Zeckhauser. The market for evaluations. *American economic review*, 89(3):564–584, 1999.

[11] Sulin Ba and Paul A Pavlou. Evidence of the effect of trust building technology in electronic markets: Price premiums and buyer behavior. *MIS quarterly*, pages 243–268, 2002.

[12] Dwight Barry. Do not use averages with likert scale data. *https://bookdown.org/Rmadillo/likert/*. Accessed: 2020-09-11.

[13] James R Bettman, Mary Frances Luce, and John W Payne. Constructive consumer choice processes. *Journal of consumer research*, 25(3):187–217, 1998.

[14] Dario Bonaretti, Marcin Łukasz Bartosiak, and Gabriele Piccoli. Cognitive anchoring of color cues on online review ratings. 2017.

[15] Nathalie Bonnardel, Annie Piolat, and Ludovic Le Bigot. The impact of colour on website appeal and users' cognitive processes. *Displays*, 32(2):69–80, 2011.

[16] Federica Cena, Cristina Gena, Pierluigi Grillo, Tsvi Kuflik, Fabiana Vernero, and Alan J Wecker. How scales influence user rating behaviour in recommender systems. *Behaviour & Information Technology*, 36(10):985–1004, 2017.

[17] Federica Cena and Fabiana Vernero. A study on user preferential choices about rating scales. *International Journal of Technology and Human Interaction (IJTHI)*, 11(1):33–54, 2015.

[18] Federica Cena, Fabiana Vernero, and Cristina Gena. Towards a customization of rating scales in adaptive systems. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 369–374. Springer, 2010.

[19] Pei-Yu Chen, Shin-yi Wu, and Jungsun Yoon. The impact of online recommendations and consumer feedback on sales. *ICIS 2004 Proceedings*, page 58, 2004.

[20] Fei-Fei Cheng, Chin-Shan Wu, and David C Yen. The effect of online store atmosphere on consumer's emotional responses–an experimental study of music and colour. *Behaviour & Information Technology*, 28(4):323–334, 2009.

[21] Alain Yee Loong Chong, Boying Li, Eric WT Ngai, Eugene Ch'ng, and Filbert Lee. Predicting online product sales via online reviews, sentiments, and promotion strategies. *International Journal of Operations & Production Management*, 2016.

[22] Jacob Cohen. Statistical power analysis for the behavioral sciences. pages 164–169, 2013.

[23] Andrew M. Colman, Claire E. Norris, and Carolyn C. Preston. Comparing rating scales of different lengths: Equivalence of scores from 5-point and 7-point scales. *Psychological Reports*, 80(2):355–362, 1997.

[24] Lluís Coromina and Germà Coenders. Reliability and validity of egocentered network data collected via web: A meta-analysis of multilevel multitrait multimethod studies. *Social networks*, 28(3):209–231, 2006.

[25] Dan Cosley, Shyong K Lam, Istvan Albert, Joseph A Konstan, and John Riedl. Is seeing believing? how recommender system interfaces affect users' opinions. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 585–592, 2003.

[26] Mick P Couper, Roger Tourangeau, and Kristin Kenyon. Picture this! exploring visual effects in web surveys. *Public Opinion Quarterly*, 68(2):255–266, 2004.

[27] Geng Cui, Hon-Kwong Lui, and Xiaoning Guo. The effect of online consumer reviews on new product sales. *International Journal of Electronic Commerce*, 17(1):39–58, 2012.

[28] Bart De Langhe, Philip M Fernbach, and Donald R Lichtenstein. Navigating by the stars: Investigating the actual and perceived validity of online user ratings. *Journal of Consumer Research*, 42(6):817–833, 2016.

[29] Anna DeCastellarnau. A classification of response scale characteristics that affect data quality: a literature review. *Quality & quantity*, 52(4):1523–1559, 2018.

[30] Dellarocas, Fan, and Wood. Reciprocity, free riding, and participation decay in online communities. pages 254–856, 2003.

[31] Chrysanthos Dellarocas. The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management science*, 49(10):1407–1424, 2003.

[32] Wenjing Duan, Bin Gu, and Andrew B. Whinston. Do online reviews matter? — an empirical investigation of panel data. *Decision Support Systems*, 45(4):1007 – 1016, 2008. Information Technology and Systems in the Internet-Era.

[33] Kenneth Fehrman and Cherie Fehrman. Color: The secret influence. pages 124–126, 2000.

[34] Ron Garland. The mid-point on a rating scale: Is it desirable. *Marketing bulletin*, 2(1):66–70, 1991.

[35] Cristina Gena, Roberto Brogi, Federica Cena, and Fabiana Vernero. The impact of rating scales on user's rating behavior. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 123–134. Springer, 2011.

[36] Paul Grice. Studies in the way of words. pages 124–126, 1989.

[37] Jiawei Han, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. pages 243–257, 2011.

[38] Saram Han and Chris K Anderson. Customer motivation and response bias in online reviews. *Cornell Hospitality Quarterly*, 61(2):142–153, 2020.

[39] Sungwon Han, Doyo Choi, and Young Cha. The effect of colour on the anchoring heuristic in consumer decision making. *Journal of European Psychology Students*, 5(3):34–36, 2014.

[40] Jon Herlocker, Joseph Konstan, Loren Terveen, John C.s Lui, and T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, 22:5–53, 01 2004.

[41] Nan Hu, Indranil Bose, Noi Sian Koh, and Ling Liu. Manipulation of online reviews: An analysis of ratings, readability, and sentiments. *Decision support systems*, 52(3):674–684, 2012.

[42] Nan Hu, Noi Sian Koh, and Srinivas K. Reddy. Ratings lead you to the product, reviews help you clinch it? the mediating role of online review sentiments on product sales. *Decision Support Systems*, 57:42 – 53, 2014.

[43] Nan Hu, Ling Liu, and Jie Jennifer Zhang. Do online reviews affect product sales? the role of reviewer characteristics and temporal effects. *Information Technology and management*, 9(3):201–214, 2008.

[44] Gerald Jacobs. Comparative color vision. page 170, 2013.

[45] Domicele Jonauskaite, Jörg Wicker, Christine Mohr, Nele Dael, Jelena Havelka, Marietta Papadatou-Pastou, Meng Zhang, and Daniel Oberfeld. A machine learning approach to quantify the specificity of colour–emotion associations and their cultural differences. *Royal Society open science*, 6(9):190–241, 2019.

[46] Raghav Pavan Karumur, Tien T. Nguyen, and Joseph A. Konstan. Exploring the value of personality in predicting rating behaviors: A study of category preferences on movielens. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, page 139–142, New York, NY, USA, 2016. Association for Computing Machinery.

[47] Claudia Kawai, Gáspár Lukács, and Ulrich Ansorge. Polarities influence implicit associations between colour and emotion. *Acta Psychologica*, 209:103–143, 2020.

[48] Cenk Kocas and Can Akkan. How trending status and online ratings affect prices of homogeneous products. *International Journal of Electronic Commerce*, 20(3):384–407, 2016.

[49] Achilleas Kostoulas. How to interpret ordinal data. *https://achilleaskostoulas.com/2014/02/23/how-to-interpret-ordinal-data/report*. Accessed: 2020-09-07.

[50] Jon A Krosnick. Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied cognitive psychology*, 5(3):213–236, 1991.

[51] Jon A Krosnick and Leandre R Fabrigar. Designing rating scales for effective measurement in surveys. *Survey measurement and process quality*, pages 141–164, 1997.

[52] Tanja Kunz. *Rating scales in web surveys. A test of new drag-and-drop rating procedures.* PhD thesis, Technische Universität, 2015.

[53] Lauren I Labrecque and George R Milne. Exciting red and competent blue: the importance of color in marketing. *Journal of the Academy of Marketing Science*, 40(5):711–727, 2012.

[54] Sandra Larrivee, Frank L Greenway, and William D Johnson. A statistical analysis of a traffic-light food rating system to promote healthy nutrition and body weight. *Journal of diabetes science and technology*, 9(6):1336–1341, 2015.

[55] Heather C Lench, Sarah A Flores, and Shane W Bench. Discrete emotions predict changes in cognition, judgment, experience, behavior, and physiology: a meta-analysis of experimental emotion elicitations. *Psychological bulletin*, 137(5):834–836, 2011.

[56] Lingfang Li. Reputation, trust, and rebates: How online auction markets can improve their feedback mechanisms. *Journal of Economics & Management Strategy*, 19(2):303–331, 2010.

[57] Xiaolin Li, Chaojiang Wu, and Feng Mai. The effect of online reviews on product sales: A joint sentiment-topic analysis. *Information & Management*, 56(2):172–184, 2019.

[58] Xinxin Li and Lorin M Hitt. Self-selection and information role of online product reviews. *Information Systems Research*, 19(4):456–474, 2008.

[59] David A. Lishner, Amy B. Cooter, and David H. Zald. Addressing measurement limitations in affective rating scales: Development of an empirical valence scale. *Cognition and Emotion*, 22(1):180–192, 2008.

[60] Michael Luca. Designing online marketplaces: Trust and reputation mechanisms. *Innovation Policy and the Economy*, 17:77–93, 2017.

[61] Warih Maharani, Dwi H Widyantoro, and Masayu L Khodra. Discovering users' perceptions on rating visualizations. In *Proceedings of the 2nd International Conference in HCI and UX Indonesia 2016*, pages 31–38, 2016.

[62] Markus A Maier, Andrew J Elliot, and Stephanie Lichtenfeld. Mediation of the negative effect of red on intellectual performance. *Personality and Social Psychology Bulletin*, 34(11):1530–1540, 2008.

[63] Bacarea V. Marusteri M. Comparing groups for statistical differences: how to choose the right statistical test?. *Biochem Med (Zagreb).*, 20(1):15–32, 2010.

[64] Kirsten Medhurst and Rashmi Sinha. Interaction design for recommender systems. *Presentation at the International Conference on Designing Interactive Systems, London, June*, pages 567–588, 03 2002.

[65] Natalja Menold, Lars Kaczmirek, Timo Lenzner, and Aleš Neusar. How do respondents attend to verbal labels in rating scales? *Field Methods*, 26(1):21–39, 2014.

[66] Wendy W Moe and David A Schweidel. Online product opinions: Incidence, evaluation, and evolution. *Marketing Science*, 31(3):372–386, 2012.

[67] Lev Muchnik, Sinan Aral, and Sean J Taylor. Social influence bias: A randomized experiment. *Science*, 341(6146):647–651, 2013.

[68] Frank Nagle and Christoph Riedl. Online word of mouth and product quality disagreement. In *Academy of management proceedings*, volume 2014, pages 156–181. Academy of Management Briarcliff Manor, NY 10510, 2014.

[69] Syavash Nobarany, Louise Oram, Vasanth Kumar Rajendran, Chi-Hsiang Chen, Joanna McGrenere, and Tamara Munzner. The design space of opinion measurement interfaces: exploring recall support for rating and ranking. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 2035–2044, 2012.

[70] Stephen M. Nowlis, Barbara E. Kahn, and Ravi Dhar. Coping with Ambivalence: The Effect of Removing a Neutral Option on Consumer Attitude and Preference Judgments. *Journal of Consumer Research*, 29(3):319–334, 12 2002.

[71] Colm O'Muircheartaigh, George Gaskell, and Daniel B Wright. Weighing anchors: Verbal and numeric labels for response scales. *Journal of Official Statistics-Stockholm*, 11:295–308, 1995.

[72] Kiemute Oyibo, Rita Orji, and Julita Vassileva. Investigation of the persuasiveness of social influence in persuasive technology and the effect of age and gender. In *PPT@ Persuasive*, pages 32–44, 2017.

[73] Paul A Pavlou and Angelika Dimoka. The nature and role of feedback text comments in online marketplaces: Implications for trust building, price premiums, and seller differentiation. *Information Systems Research*, 17(4):392–414, 2006.

[74] John W Payne, John William Payne, James R Bettman, and Eric J Johnson. The adaptive decision maker. pages 34–36, 1993.

[75] Richard E Petty and John T Cacioppo. The elaboration likelihood model of persuasion. In *Communication and persuasion*, pages 1–24. Springer, 1986.

[76] Hau Xuan Pham and Jason J Jung. Preference-based user rating correction process for interactive recommendation systems. *Multimedia tools and applications*, 65(1):119–132, 2013.

[77] Olga Pilipczuk and Galina Cariowa. Opinion acquisition: An experiment on numeric, linguistic and color coded rating scale comparison. In *International Multi-Conference on Advanced Computer Systems*, pages 27–36. Springer, 2016.

[78] Paul Resnick and Richard Zeckhauser. Trust among strangers in internet transactions: Empirical analysis of ebay's reputation system. *The Economics of the Internet and E-commerce*, 11(2):23–25, 2002.

[79] Christoph Riedl, Ivo Blohm, Jan Marco Leimeister, and Helmut Krcmar. The effect of rating scales on decision quality and user attitudes in online innovation communities. *International Journal of Electronic Commerce*, 17(3):7–36, 2013.

[80] Matthew J Salganik, Peter Sheridan Dodds, and Duncan J Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *science*, 311(5762):854–856, 2006.

[81] Ann E Schlosser. Can including pros and cons increase the helpfulness and persuasiveness of online reviews? the interactive effects of ratings and arguments. *Journal of Consumer Psychology*, 21(3):226–239, 2011.

[82] Norbert Schwarz, Carla E Grayson, and Bärbel Knäuper. Formal features of rating scales and the interpretation of question meaning. *International Journal of Public Opinion Research*, pages 44–56, 1998.

[83] Anuj K Shah and Daniel M Oppenheimer. Heuristics made easy: An effort-reduction framework. *Psychological bulletin*, 134(2):207–209, 2008.

[84] Cristina Soriano and Javier Valenzuela. Emotion and colour across languages: implicit associations in spanish colour terms. *Social Science Information*, 48(3):421–445, 2009.

[85] E Isaac Sparling and Shilad Sen. Rating: how difficult is it? In *Proceedings of the fifth ACM conference on Recommender systems*, pages 149–156, 2011.

[86] Shrihari Sridhar and Raji Srinivasan. Social influence effects in online product ratings. *Journal of Marketing*, 76(5):70–88, 2012.

[87] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. Introduction to data mining. pages 144–156, 2016.

[88] Vera Toepoel and Don A Dillman. Words, numbers, and visual heuristics in web surveys: Is there a hierarchy of importance? *Social Science Computer Review*, 29(2):193–207, 2011.

[89] Vera Toepoel, Brenda Vermeeren, and Baran Metin. Smileys, stars, hearts, buttons, tiles or grids: influence of response format on substantive response, questionnaire experience and response time. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, 142(1):57–74, 2019.

[90] Roger Tourangeau, Mick P Couper, and Frederick Conrad. Color, labels, and interpretive heuristics for response scales. *Public Opinion Quarterly*, 71(1):91–112, 2007.

[91] Roger Tourangeau, Lance J Rips, and Kenneth Rasinski. The psychology of survey response. pages 222–237, 2000.

[92] Dimitrios Tsekouras. The effect of rating scale design on extreme response tendency in consumer product ratings. *International Journal of Electronic Commerce*, 21(2):270–296, 2017.

[93] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *science*, 185(4157):1124–1131, 1974.

[94] Jeroen Van Barneveld and Mark Van Setten. Designing usable interfaces for tv recommender systems. In *Personalized Digital Television*, pages 259–285. Springer, 2004.

[95] Kevin E Voss, Eric R Spangenberg, and Bianca Grohmann. Measuring the hedonic and utilitarian dimensions of consumer attitude. *Journal of marketing research*, 40(3):310–320, 2003.

[96] Eric Walden. Some value propositions of online communities. *Electronic Markets*, 10(4):244–249, 2000.

[97] Bert Weijters, Elke Cabooter, and Niels Schillewaert. The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing*, 27(3):236–247, 2010.

[98] Bert Weijters, Maggie Geuens, and Hans Baumgartner. The effect of familiarity with the response category labels on item response to likert scales. *Journal of Consumer Research*, 40(2):368–381, 2013.

[99] Bert Weijters, Maggie Geuens, and Hans Baumgartner. The effect of familiarity with the response category labels on item response to likert scales. *Journal of Consumer Research*, 40(2):368–381, 2013.

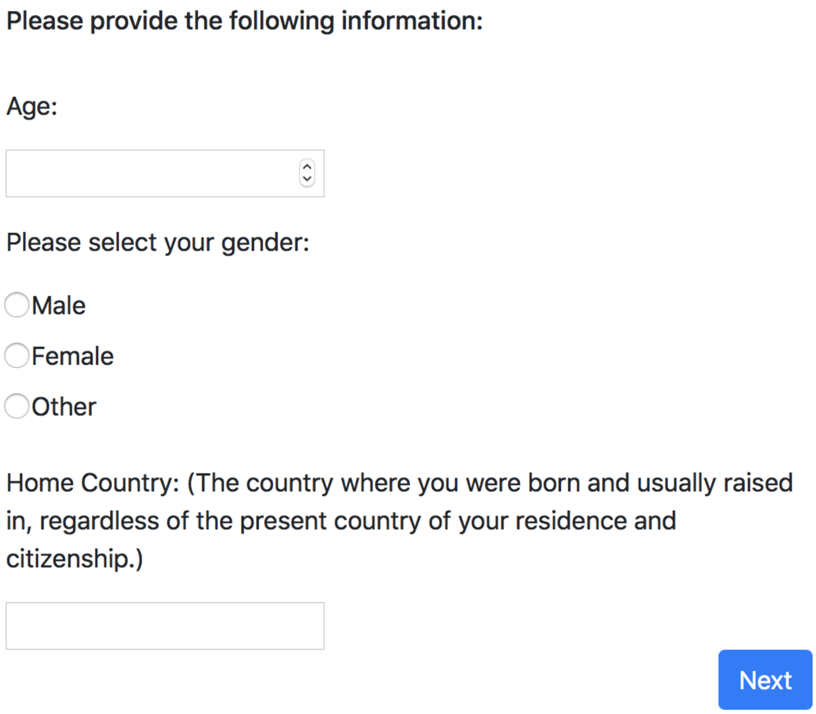
[100] Albert R Wildt and Michael B Mazis. Determinants of scale response: Label versus position. *Journal of Marketing Research*, 15(2):261–267, 1978.

[101] Ming Zhou, Martin Dresner, and Robert Windle. Revisiting feedback systems: Trust building in digital markets. *Information Management*, 46(5):279 – 284, 2009.

Appendix A

USER STUDY

A.1 User Demographic



Please provide the following information:

Age:

Please select your gender:

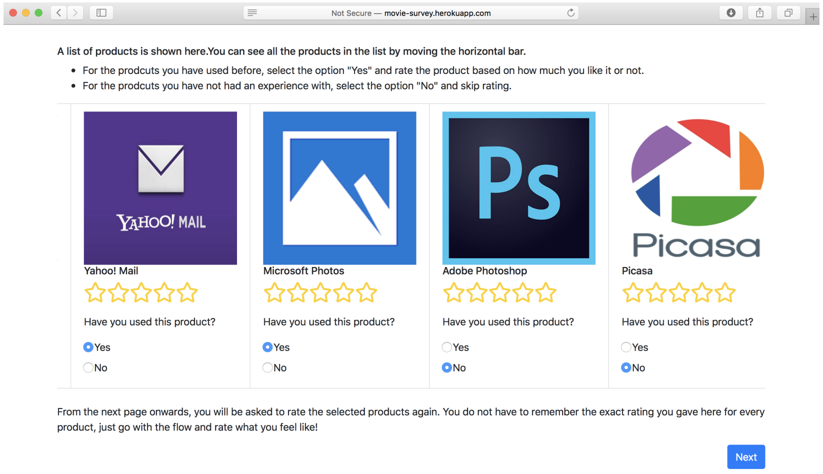
○ Male

○ Female

○ Other

Home Country: (The country where you were born and usually raised in, regardless of the present country of your residence and citizenship.)

A.2 Product Selection and Rating



A list of products is shown here.You can see all the products in the list by moving the horizontal bar.

- For the prodcuts you have used before, select the option "Yes" and rate the product based on how much you like it or not.
- For the prodcuts you have not had an experience with, select the option "No" and skip rating.

Yahoo! Mail
☆☆☆☆☆
Have you used this product?
● Yes
○ No

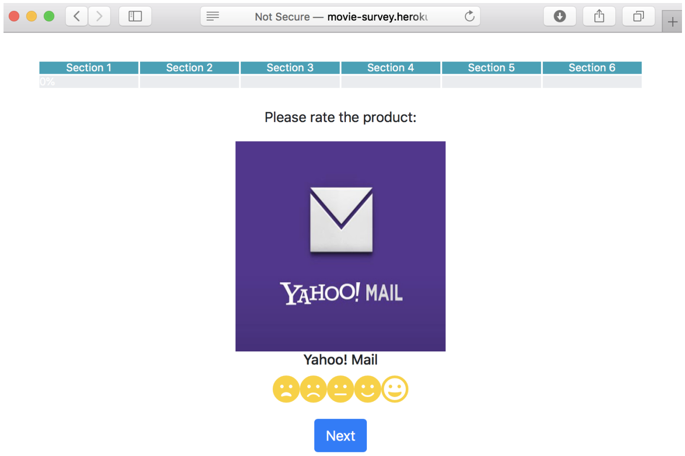
Microsoft Photos
☆☆☆☆☆
Have you used this product?
● Yes
○ No

Adobe Photoshop
☆☆☆☆☆
Have you used this product?
○ Yes
● No

Picasa
☆☆☆☆☆
Have you used this product?
○ Yes
● No

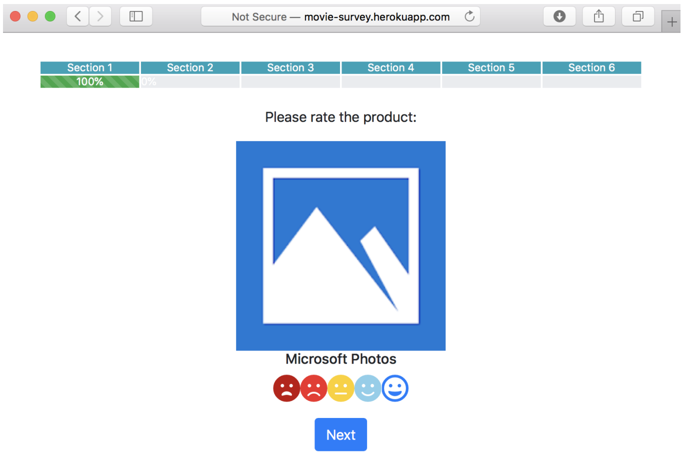
From the next page onwards, you will be asked to rate the selected products again. You do not have to remember the exact rating you gave here for every product, just go with the flow and rate what you feel like!

A.3 Rating Selected Product with Six Scales



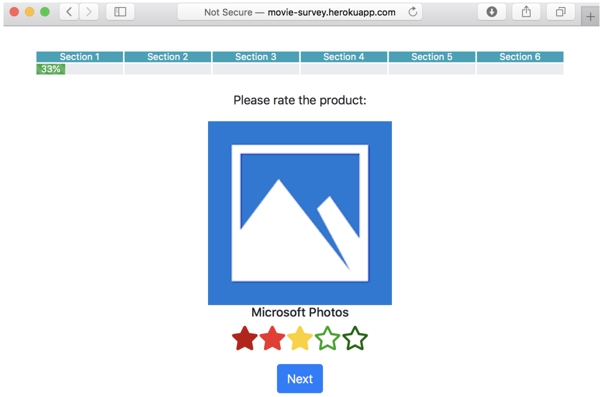
Please rate the product:

Yahoo! Mail



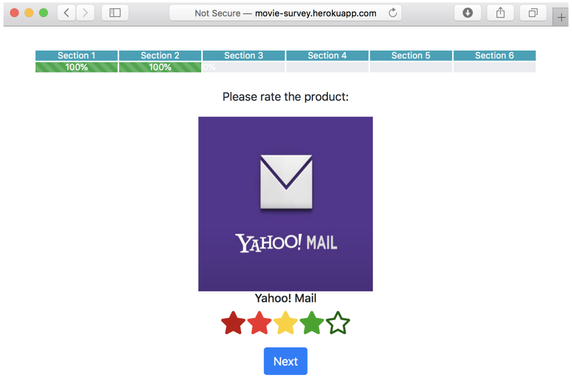
Please rate the product:

Microsoft Photos



Please rate the product:

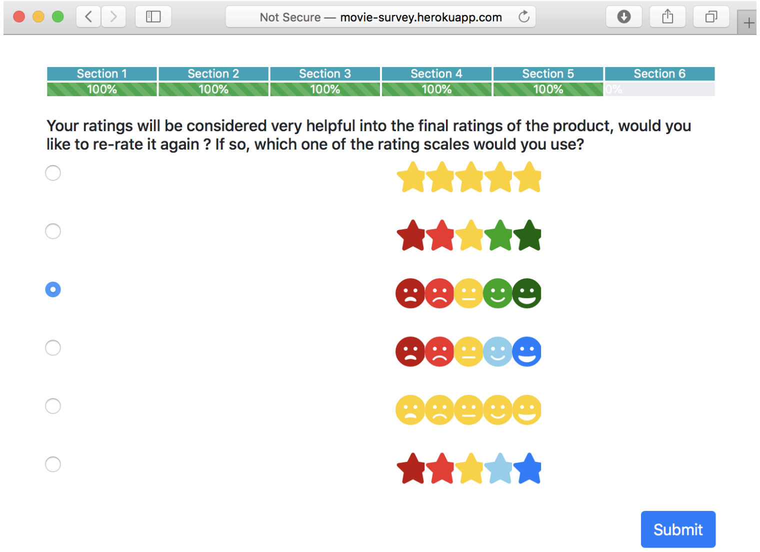
Microsoft Photos



Please rate the product:

Yahoo! Mail

A.4 Preefrred and Common Scale



Section 1	Section 2	Section 3	Section 4	Section 5	Section 6
100%	100%	100%	100%	100%	0%

Your ratings will be considered very helpful into the final ratings of the product, would you like to re-rate it again ? If so, which one of the rating scales would you use?

○ ★★★★★

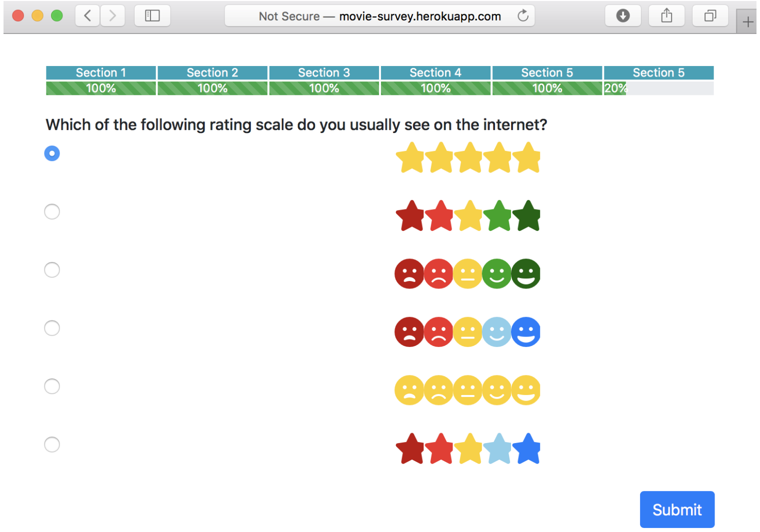
○ ★★★★★

● ☹☹😐☺☺

○ ☹☹😐☺☺

○ ☹☹😐☺☺

○ ★★★★★



Section 1	Section 2	Section 3	Section 4	Section 5	Section 5
100%	100%	100%	100%	100%	20%

Which of the following rating scale do you usually see on the internet?

● ★★★★★

○ ★★★★★

○ ☹☹😐☺☺

○ ☹☹😐☺☺

○ ☹☹😐☺☺

○ ★★★★★