Collaborative reflection on personal data: An approach for investigating context-related user experiences in recommender systems

Version of September 18, 2024

Zihan Wang

Collaborative reflection on personal data: An approach for investigating context-related user experiences in recommender systems

THESIS

submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

in

COMPUTER SCIENCE

by

Zihan Wang

TU Delft

Crowd Computing and Human-Centered AI Group
Department of Web Information Systems
Faculty EEMCS, Delft University of Technology
Delft, the Netherlands
www.ewi.tudelft.nl

Collaborative reflection on personal data: An approach for investigating context-related user experiences in recommender systems

Author: Zihan Wang
Student id: 4840438

Abstract

Recommender systems are widely used in modern lives and contribute to many industries. Therefore, methods to evaluate and improve them are important. Nowadays, much research has been done to improve the system aspects such as algorithms. However, user experiences are not only affected by the systems but heavily rely on the context when using the systems. Therefore, the research on user aspects to understand their experiences is as important. This study contributes an approach that uses collaborative reflection to find insights into users' experiences with recommender systems. Using this approach, this study presents the influences of context on user experiences with recommender systems. This study investigates the importance of situational and personal contexts like mood, time, and location in shaping user satisfaction with recommendations. The research adopts a method based on collaborative reflection, where participants engage in tasks using their YouTube watch history, paired with another individual for real-time discussion. By analyzing contextual influences and the values users wish to achieve, the study identifies key patterns in user behavior and insights into personal preferences. This research not only contributes to the evaluation of recommender systems but also highlights the need for systems to align with both the goals of users and broader societal values. The usability of the proposed method was tested to be successful in crowdsourcing, yielding practical implications for future evaluation and improvements of recommender systems.

Thesis Committee:

Chair:	Qing Wang, Faculty EEMCS, TU Delft
University supervisor:	Jie Yang, Faculty EEMCS, TU Delft
Daily supervisor:	Di Yan, Faculty IDE, TU Delft

Preface

I would like to express my deepest gratitude to everyone who supported and guided me throughout this thesis project. First and foremost, I am incredibly thankful to my academic supervisors, for their invaluable advice, insightful feedback, and unwavering support during the entire process. Their expertise and encouragement pushed me to explore new ideas and continually improve my work. My appreciation extends to all the participants in my experiments, whose willingness to engage and offer their time was essential for this study. Their contributions provided the rich data that forms the foundation of this research.

Lastly, I would like to thank my family and friends for their constant encouragement and patience, especially during the most challenging stages of this journey. Your unwavering belief in me has been a source of strength throughout this process.

Thank you all for making this thesis possible.

Zihan Wang
Delft, the Netherlands
September 18, 2024

Contents

List of Figures

Chapter 1

Introduction

Recommender systems have been widely used to provide users with targeted content to improve user satisfaction. It is adopted in multiple areas such as video platforms, e-commerce, and social media. Using recommendation systems in online shopping improves sales and increases flexibility for retailers to adjust their prices[7]. R. Zhou[54] and his team found a strong correlation between a video's view count and the average view count of its top referrer videos, proving the influence of recommendations. Even if much research has shown the benefits of recommender systems, the quality of recommendations still has the potential to be improved. Enhancing user satisfaction with recommended items remains a topic of ongoing discussion.

The recommender systems now usually make use of either or both collaborative filtering(CF) and Content-Based Filtering(CBF)[1]. According to a user-centric evaluation framework proposed by Bart P. Knijnenburg et al, users' experiences are affected by system aspects, situational characteristics, and personal characteristics[29]. The two methods we mentioned above are improvements from systems aspects. However, exploring possible improvements from the user side(situational and personal characteristics) is as important. In real life, users' experiences can be influenced by many factors, for example, moods, time, location etc. A change in any of these can affect the users' final decision. One research by R. Logesh and V. Subramaniyaswamy proposed a method to recommend travel locations by taking contexts, such as locations, weather, and past travel history, into consideration[34]. Their results show that incorporating users' contextual information is helpful and justifies the importance of finding richer insights into users' attitudes toward an item. Investigation into these aspects will benefit evaluation and modelling by understanding how contexts influence user experiences.

Besides the need to understand the effects of context, making the recommendations align with human values is just as important. As AI techniques develop, how to make the results computed by AI align with human value has become a hot topic. According to Shalom H. Schwartz, values refer to people's desired goals and guide the selection or evaluation of actions, and policies[44]. How much people favour the recommendations given is influenced by human values and values guide their decision-making[41]. Making the recommendations align with the goals of users, system designers, and society should be one of the objectives that any recommender system should pursue[47]. In the recommender systems context, values should be extended and clarified. Besides classical human values, such as fairness and diversity, Jonathan Stray et al listed the values relevant to recommender systems[47]. There is little research on under what situations users will require these values in the recommender systems context and the process of eliciting values from users is complex.

Nowadays, only a few recommender systems have deployed measures to collect user feedback to understand why they like/dislike a recommended item but they are limited. YouTube and Netflix only provide choices of likes or dislikes. TikTok provides a relatively wider range of options but is still limited to reasons about video content. To fill the gap and collect deeper insights into users' experiences when using recommender systems, this research aims to provide a method to investigate how context influences user experiences on recommender systems. Among existing research on helping people find inner thoughts corresponding to actions in psychology and design, reflection is considered one of the most effective ways to enable individuals to generate insights for understanding themselves and self-improvement[10][5]. Research by Y.Tawatsuji et al explained how humans improve their problem-solving abilities by learning from mistakes. Therefore, making participants reflect on their experiences is regarded as the foundation of our method. The kind of reflection we selected is collaborative reflection. Unlike personal reflection where people check their data and answer some questions, collaborative reflection enhances participants' understanding of each other's data and creates connections. It helps participants gain insights by linking or comparing their own experiences to others' reflections on their own experiences. In the process, we expect them to generate insights that are useful to improve recommendation quality during discussions.[21][39]. However, effective collaborative reflection builds on powerful support for the communication between participants. Therefore, how to use collaborative reflection in this setting is one of the topics the research cares. In 2017, E.Choe et al made research on how people how people reflect on personal data through visual data exploration[11]. They found several levels of reflection that users get from the provided visualizations. Their research supports the effectiveness of using data visualization to help people reflect.

Our method adopts the points mentioned above to help reflect: Providing data visualizations and prompting collaborative reflection. Based on them, we designed a platform where participants worked on tasks related to their YouTube watching habits. We selected YouTube in the case study because of its tremendous amount of users and its convenient API[20]. It made both platform building and participant recruitment easy. On our platform, two participants are paired automatically and finish interactive tasks together in a video chat. We help participants reflect on insights into why they like/dislike an item recommended to them by providing visualizations of their YouTube data to check and operate. To make this method easier for recommender system providers to use and to cover diverse user groups, we wish our method to be usable for crowdsourcing. Therefore, we deployed our platform on Prolific, a crowdsourcing platform. Crowdsourcing is considered an effective way to gather data from people[23]. According to Aniket Kittur et al, crowdsourcing can lead to satisfying answers for subjective questions[28]. The tasks mentioned in their research and most crowdsourcing tasks nowadays are individual micro-tasks. Although there is research on facilitating cooperative crowdsourcing such as crowd4U[25] and turkomatic[30], they achieved the target by splitting complex tasks into sub-tasks and assigning them to crowd workers to form a pipeline. Almost no research has been done on collaborative crowdsourcing that requires two crowd workers to cooperate on a task in real time. The challenges and limitations we met during the process are summarized and expected to contribute to future research in designing online real-time collaborative crowdsourcing tasks.

1.1 Research Question(s)

In conclusion, understanding how contexts influence user experiences is crucial. It is also important to make the recommendations align with the values of each party. However, few explorations have been done to explore the most efficient way to do it. Therefore, our research aims to fill in the gap of using collaborative reflection on crowdsourcing to study users of recommender systems and proposes the following research questions:

- How should we take context into consideration while evaluating recommender systems?
- How can online collaborative reflection on personal data help find insights into user experiences on recommender systems?

Chapter 2

Related Work

Ways of evaluating users' experiences with recommender systems have been explored extensively. This chapter discusses state-of-arts related to our target of using collaborative reflection on crowdsourcing platforms to study context-related user experiences of recommender systems. This chapter is divided into four sections. In the first section, we will talk about existing approaches to user study, In the second section, we will introduce collaborative reflection and ways to support it, followed by our review of value alignment, especially about existing works to make recommendations align with human values. The last section will introduce the existing recommendation algorithms used by recommender systems to present why we need improvements from user aspects.

2.1 User study for recommender systems

According to the user-centric evaluation model of recommender systems proposed by Knijnenburg et al, situational characteristics and personal characteristics are as important as system aspects in improving user experiences[29]. Situational characteristics are often dependent on the context when interacting with the system. Personal Characteristics mainly include users' demographics and cognitions. Investigating the user sides of recommender systems is also important as we have introduced the necessity of holistic modeling and value alignment. However, not much research has been done in this area. The most direct way to study users is by gathering their feedback. User feedback can be divided into two categories: Explicit feedback and implicit feedback. Most recommender systems research focuses on improving from explicit feedback rather than implicit user feedback[24], but there is evidence that a combination of implicit and explicit user feedback will improve the performance of recommender systems[26]. It again reflects the importance of research on user sides for recommender systems.

Crowdsourcing as a powerful method to gather data from crowds, can be used in user studies[28]. In 2012, Marion K. Poetz and Martin Schreier investigated new product ideas from crowdsourcing and compared them with those thought of by experts. It turns out user ideas score higher on novelty and customer benefit, although lower in feasibility[38]. Their results implied that users are better at expressing their needs than technology-related problems. Tara S. Behrend et al examined the viability of crowdsourcing for survey research and initially proved that the data quality is as good as that from regular pools[4]. Their results provide theoretical support for using crowdsourcing platforms to do user studies.

2.1.1 Reflection on personal data

Around 2010, the rise of personal informatics made the first step to providing people a chance to view and reflect on their data. This method can be used in this research to trigger participants' thinking in the context when they watch a video. As subsequent studies continued to explore, methods were developed to guide users to gain an understanding of themselves. Ian Li et al found that tools designed at that time did not sufficiently understand users' self-reflection needs[32]. Therefore, they conducted a study and identified six kinds of questions people care about their data: Status, History, Goals, Discrepancies, Context, and Factors. In the six perspectives, they summarized what participants were interested in:

- Status: Data that revealed their current status.
- History: Trends and patterns over the long term.
- Goal: Goals that are appropriate to pursue
- Discrepancies: How is their current status different from their goal?
- Context: What was happening at or near the same time as the context they are seeking?
- Factors: How various factors affected the current status.

These points guided this research on task design. They are regarded as points that can encourage users to talk about their feelings.

In 2017, Choe et al investigated how people reflect on their data through visual data exploration. Their research helps us understand how people interact with their data and the insights they get along the process. The research shows that people are excited about understanding themselves through visual data exploration. They found that peaks and extreme values caught participants' eyes and comparing various periods promoted people to recall their past behaviors. Also, they found recalling past behaviors brings out the users' curiosity, which leads them to continue to explore.[11]. Another research by Alice et al investigated how the construction of personal physicalizations helps people reflect. They found that people gained insights from levels of depth from simply reading data to contemplations of personal values and attitudes[50]. The levels are Reflections on Data, Context, Action, and Values. These findings are valuable to our research because we want to use the same method to figure out what influences people's attitudes toward recommended items. The reflections from different levels are valuable. The values that users want to get while using a recommender system are something that we should also consider when designing algorithms. Besides what's mentioned above, J.Cho et al[10] identified shortcomings of the ways that personal informatics use personal reflection and provide guidance on how to support reflection when designing such a system. A method called ”Guess the Data” easily persuades people to answer questions that are not easily answerable, and it stimulates curiosity by letting users take a guess and then let them explore the real data[31].

2.2 Collaborative Sensemaking

Since a model that can capture the key attributes of users and items is essential for the algorithms, figuring out what attributes should be considered when giving recommendations is an indispensable first step. Instead of only considering items' attributes, we aim to dig deeper into the insights that could influence users's preferences. This includes the contexts while using the system and the values they wish to receive.

Russell's team defined sensemaking as ”the process of searching for a representation and encoding data in that representation to answer task-specific questions”[42]. It can also be done by multiple people, which is collaborative sensemaking. It is a sensemaking process involving people with different viewpoints/backgrounds. It is proved that pair work facilitates plausibility growth and thus each individual's sensemaking[52]. People can exchange their ideas and give feedback to each other in collaborative sensemaking and this supports people set up connections between data and environmental factors[40]. This helps them reframe their understanding of their data as they get the opportunity to see their data from others' perspectives. On the other hand, they can compare each other, potentially bringing insights. Accession of another person makes the process explicit and engaging[22]. Research by Naomi F Dale et al used paired interviews to get a greater understanding of the interactions between tourism and hospitality service participants. The idea of focusing on interactions between participants instead of only the work they finished is valuable to this study[12].

To make participants generate collaborative sensemaking, providing support during the process is necessary. Different from reflecting on your personal data, collaborative sensemaking requires participants to make their own experiences explicit and understandable to others. Only in this way, can they learn and compare them, and finally get some insights on the reasons behind the experiences and propose changes to make[39]. It has been proved that structuring the reflection process is helpful in face-to-face scenarios[13]. It means that the setup should make the reflection target clear to the participants and provide anchors that participants can reflect on[14]. According to Michael and Bettina's summary of aspects of collaborative reflection from multiple research, the most recurring topics are[39]:

- ”Recall of past experiences
- Probing, challenging or supporting other contributions
- Linking experiences with other sources(experiences, knowledge, data)
- Repeating and commenting on other contributions
- Sharing opinions
- Drawing from experiences

- Transforming insights into practice”

Based on the idea of promoting this, they built a prototype app to test their tools for promoting collaborative reflection and summarized the roles of tools in collaborative reflection. They mentioned that the tool should make participants engaged in reflection for a period of time, strengthen the connection between participants, and ”Balance diversity and intensity of communication”. As we mentioned in 2.3.1 about the reflection of personal data, providing people with their own data can help them reflect on their behaviours. The same goes for collaborative sensemaking. Mahyar and Tory investigated how demonstrating recorded information supports communication and coordination in collaborative sensemaking. The result showed that by providing recorded information, groups could communicate more effectively and coordinate better, which leads to better results[35].

2.3 Value alignment

With the development of AI technology, the alignment problem becomes an important topic to be worked on. A properly aligned AI should take all unethical or imprudent behaviors into account and prevent them[18]. As for recommender systems, the alignment problem becomes how should we recommend items such that they can serve various human values. Jonathan Stray et al proposed a design idea that recommender systems should align recommendations with the goals/purposes of users, system designers, and whole society[47]. They also mentioned that the goals for using recommender systems are often context-dependent and consequential. Therefore, incorporating value-based labels when modeling items and users in recommender systems is necessary.

Value-sensitive design is an approach that ”accounts for human values in a principled and comprehensive manner throughout the design process[17].” This approach should be adopted when designing a recommender system such that the recommendations can align with the values we wish to receive from it. In the value-sensitive design area, values are defined as “what a person or group of people consider important in life”[8]. When we put values in recommender systems, it is important to know which human values are applicable. Research by Jonathan et al[47] collected values that are particularly relevant in recommender contexts and they helped in our analysis.

2.4 Recommendation algorithms

Content-Based Filtering and Collaborative Filtering are the two most common methods used nowadays. In the introduction, we have given a summary of their specialities, and we will provide a more detailed description of them to present the limitations of relying only on improvements in system perspectives.

One drawback of Content-Based Filtering(CBF) is its limited potential to capture complex user actions. First of all, CBF models both users and items. CBF models users and items by their attributes. This leads to the problem that the accuracy of recommendations using CBF completely depends on how accurately we model items and users and the influence of people’s interactions is ignored[16].

Collaborative Filtering(CF) is subdivided into user-based and item-based. In user-based CF, the user is recommended items based on the ratings given by neighbors who share the same tastes[43]. Item-based CF is recommended based on user-item interactions. Similar items will be recommended if a user has ever shown an interest in one item. This method is considered more accurate than user-based filtering because the items to be evaluated are large while users may only be interested in a tiny part of it[49]. However, it suffers from the cold start problem.

To make up for the shortcomings of CF and CBF, the idea of unifying CF and CBF was brought up. A hybrid approach can usually achieve higher accuracy in recommendations compared to CF and CBF. This is because CF lacks the information about domain dependencies and CBF ignores how people influence each other[19].

However, it still requires an accurate model that can include the necessary attributes of both items and users. To better satisfy users' requirements, the investigation into the personal and situational context is indispensable.

Chapter 3

Methods

This study employs qualitative research to find the insights that influence perceptions towards recommendations given by recommender systems. Using the idea of self-reflection from personal data and collaborative sensemaking, and YouTube as a case study, we set up a website where participants are paired and then cooperate on the reflective tasks. We publish our tasks on Prolific, a crowdsourcing platform. This section will introduce our participants, design principles, task descriptions, data collection, and analysis.

3.1 Design Principles

As we have introduced in the introduction, the goal we want to achieve while with the tasks is to facilitate participants to reflect on the reasons they watched some videos and how they are related to their values. We expected participants to talk about the insights they decided to watch a video in a relaxing discussion. According to Alice Thudt et al's research[51], there are four types of reflections: Reflections on data, context, action, and values. With these goals in mind, we did a further literature review on how to encourage users to reflect on themselves and summarized the following three design principles:

- P1: Providing the foundation for reflection by co-construction to facilitate recall of past experiences/contexts
- P2: Encouraging communication upon cooperating for an answer to link and merge experiences.
- P3: Prompting contextual judgements of experiences to elicit personal values.

3.1.1 Providing the foundation for reflection by co-construction to facilitate recall of past experiences/contexts.

The reason that we came up with this design principle can be explained from two perspectives: Helping participants to recall their experiences and facilitating collaborative reflection.

According to research by Smith et al[46], memory processes are influenced by environmental context. The contextual information, in turn, influences memory effects. Therefore, helping participants recall the context of their past experiences is a crucial step. When participants review the visualizations generated from their data, they are expected to find patterns. Combining these patterns with the information demonstrated in the visualizations, we wish participants the ability to think about the reasons behind these patterns and their feelings towards a video at that time. Contextual information acts as the trigger for participants to recall memories and reason why patterns emerge.
On the other hand, making participants familiar with their works and providing possibilities to refer to each other is the foundation for effective collaborative reflection.[39] Instead of showing the visualizations directly, we let the participants cooperate to create a visualization with data from both. In this way, participants not only get to know their own data by interacting with the visualizations but also know each other's data better since they work in the same area. Since participants have to slow down to log their data and get familiar with them in the process, participants can better understand their data as reflections occur not only from completed visualization but also during the preparation phase such as data logging, token preparations, and token integration[51]

3.1.2 Encouraging communication upon cooperating for an answer to link and merge experiences

The second principle aims to prompt participants to link their experiences with others to find new insights eventually. To achieve this, we support participants in recording their findings and data, which is proven to help the team organize and share their results[35]. Also, we designed questions that can only be answered by combining their data. We expect this idea can help participants understand their needs better and eventually reflect on their experiences while using YouTube[6]. In our tasks, we asked the two participants to work on the same chart and came up with questions that one person led the other to check each other's data. In this way, we expected them to find insights by comparing with each other, explaining their own data visualization and expanding further discussions from which.

3.1.3 Prompting contextual judgements of experiences to elicit personal value.

The third value came up to help us elicit related values while users are using recommender systems. It was found by previous research that respondents engaged in high-level reflection less frequently than "lower" level reflection[51]. Reflection on values is undoubtedly a high-level reflection. Therefore, measures have been taken to provide participants with a reasonable workflow to facilitate their thinking on values instead of asking questions about values without any padding. This process makes participants generate lower-level reflections such as reflection on the data itself first, which are the indispensable basis for generating high-level reflections such as personal values[51].

According to Pieter Desmet, emotions play an important role in eliciting values because we are only emotional about things that touch our moral or personal values[15]. We deduced it is easier to elicit human values from participants' discussions when they notice their emotion changes. Therefore, we asked participants to judge how they were satisfied with the videos they watched, which is expected to trigger their memories of the emotions after watching the videos. Combining with the contextual information we provided in the visualizations and what they recalled from previous steps, we hope participants can generate reflections on values they wished to be fulfilled while using YouTube at that time.

3.2 Task Descriptions

The complete user journey can be roughly divided into two phases: Preparation and cooperative tasks. In the preparation phase, participants download their YouTube watch history from Google Takeout and wait to be paired in a waiting room. As soon as we have two participants in the queue, we direct them to a room where they cooperate on their tasks. All the tasks are designed to minimize the number of operations participants need so that they can focus on discussions.

3.2.1 Preparation

The first page participants see is a general introduction to the tasks they will work on. It tells them that our tasks can help them reflect on their own YouTube-watching habits so that they can get to know themselves better. This is considered a measure to get better data quality from crowd workers by motivating them. The method of getting reliable crowdsourcing data is summarized from multiple research by Jorge Carvalho et al in their review on improving data quality by enabling positive user experience[9]. Besides the external motivation(money reward), we expect to provide participants with internal motivation to understand themselves better to improve data quality[33]. After participants give their consent, they will be guided to download their own YouTube watch history from Google Takeout. The exportation of data from Google takes five to ten minutes, and they will be directed to the waiting room to be paired with another participant while waiting for the data. After two participants are successfully paired, they are directed to the workspace where they cooperate. They were required to record the workspace tab and join the video chat with their teammate.

3.2.2 Task design

The tasks start with a simple ice-breaking phase. We get our qualitative results from the participants' discussions, so it is important to facilitate communication between them. As we have mentioned in the participants selection section, we hire participants sharing the same backgrounds in each batch to facilitate communication. This ice-breaking phase and the requirements to be in a video chat, instead of a voice chat also contribute to facilitating communication. Participants will do a self-introduction first, followed by asking each other an ice-breaking question we provided. We also guide how to use the Miro board in this stage so that their enthusiasm for communication is not extinguished by trying to figure out how to operate the items on the board.

The remaining two tasks were designed by the principles we introduced before. To make it more clear, we listed how each task aligns with the design principles. We also pointed it out in the appendix showing directly how principles and questions correspond to each other.

- P1: all discussions while working on visualizations, 2.1.1(Stage 2, question 1, sub-question 1), 2.1.2, 2.2.2, 3.1.1, 3.2.1
- P2: 2.2.1, 3.3.1, 3.3.2
- P3: 2.2.3, 3.1.2, 3.1.3, 3.2.2, 3.2.3

The second task focuses on reflecting on the correlation between time and the videos watched. Participants will be asked to estimate the average number of videos they watch during each specified time period daily. The time periods are midnight(00:00 - 06:00), morning(06:00 - 12:00), afternoon(12:00 - 18:00), around Dinner(18:00 - 21:00), and late night(21:00 - 24:00). The evening hours are divided more carefully because this is the time when most people have free time to watch YouTube. After guessing, they can see the bar chart automatically generated from watch history. By comparing the chart of their guessing and real data, two participants are asked to discuss the following ideas. The complete questions are in Appendix A.

- How they made a guess.
- Comparison between the guess and real data
- Reflection on data patterns and personal context

The logic of question design here is first to make participants notice the pattern that may have special meanings, which is a reflection of the data itself. From this, we guide them to think about the reasons leading to these patterns and hopefully some insights related to their personal value. We also set questions that can help promote collaborative sensemaking as they need to explain their own data to their teammate and reflect on each other's data. Only by comparing and listening to the other's explanation, can they get the answer. The process forces them to recall their YouTube-watching habits and life patterns. We assume that people use YouTube for different purposes at different times. From this task, we expect to get insights into how time influences their YouTube use such that these attributes can be considered in recommendations. In this way, the contents recommended can be adjusted according to time.

The third task, which is also the last one, focuses on looking back at what happened on one day. Focusing on one day makes it possible to demonstrate enough details for participants to recall the details that happened on that day. Different from reflecting on the visualization of a week's data in stage 2, this stage aims to support participants to reflect on specific videos. As shown in the picture below, the videos participants watched yesterday are automatically generated in the chart, and sorted into the same periods as stage 2. The first step for participants is to recall what they did yesterday and fill them in the purple rectangles. The finished visualizations represent the chronological order of activities and the videos they watched and it is expected to provide details for participants to recall the context when they watched those videos. Based on the information, participants are then asked to freely choose 5 of their videos and evaluate how each satisfied them. By this point, the participant should have already recalled a preliminary structure of what happened yesterday. Based on this, they are asked questions from the following perspectives:

- The correlation between what they did and the videos they watched
- The purpose for watching videos
- Reasoning behind their most satisfying and disappointing video
- Recommend one video that would interest their teammate to promote collaborative sensemaking.

The complete questions are appended in Appendix A. We first ask participants to choose the most satisfying and the most disappointing video, and we expect them to reflect on both the context and the video itself. While they skim over the videos, the context information such as activities and time helps them form the reflection on context. They are also asked about how the videos satisfied/dissatisfied them and how is this evaluation related to their value. In the end, they are asked to recommend one video that their teammate would be interested in from the videos they watched. This is a method to facilitate collaborative reflection. We expect participants to review each other's data and give recommendations based on their favourites. It is hoped that participants will gain an understanding of how others perceive their viewing behaviour and help them to reflect from the perspective of others.

3.3 Implementations

We hosted a website where two participants could prepare their data and then pair with another participant to reflect on their YouTube watch history. Miro board [37] is embedded in our platform as the workspace where participants cooperate. We have a back-end server written by Node.js. We set up several endpoints to handle the logic of pairing participants, processing YouTube history uploaded by participants, and generating visualizations. The server processes the history file and then sends requests to Miro API to generate visualizations. It is also where we store data and recordings of the participants. The front end was built with Vue3. It takes input from users and sends requests to the endpoints on our back-end server.

A difficulty we met in implementations is how to make the platform serve well on crowdsourcing platforms. Although we tried proving crowd workers more motivations besides money by telling them that they could better understand their YouTube-watching habits during the process, our pilot tests did not go as smoothly as we expected. A significant number of participants joined the study but did nothing. As we need to pair two participants together in the study and we can only open limited seats, their behaviour of occupying positions but not working affected efficiency. To solve this problem and reduce participants' waiting time, we first added an attention check on every page and a participant's submission will be returned(Leaving the study without getting a reward). Another measure we took was to open up twice the positions we needed. In this way, we made sure that there were always candidates joining the queue to pair with another participant, which significantly reduced the number of participants who left because of tired of waiting. Another stage that participants left was after they were paired and saw the tasks on the Miro board. To handle the situation one participant quits and leaves the other one alone in the room. We implemented a feature that can directly bring one participant from the waiting room. If no one is in the waiting room at the moment, the next one who joins the queue will be directly leaded to this room. To reduce the number of participants who leave at this stage, we found another selection criterion on Prolific: Willing to join a video interview. After setting this filter, the percentage of people who left at this stage decreased.

3.4 Participants

The participants were hired both online and offline, and most were hired from the online crowdsourcing platform, prolific. It provides the functionality to select only eligible participants. According to Van Der Wege, M et al, people's communications are based on consumption they share a similar background to a certain extent. When this consumption does not hold, the communication can be impeded[53]. Also, it is easier for people to get familiar with each other if they share the same backgrounds[2]. Under this principle, we set up two groups of participants. The list below details the attributes of each group.

- European countries citizens aged from 20-35, any gender, actively using YouTube, playing video games, and willing to take part in a video interview(13 groups).
- Master students, any gender, actively using YouTube(4 groups)

The age range makes sure participants do not have too many differences with each other while ensuring we have enough candidates in the participants pool. Since our platform embeds Miro as the interactive board, we wish the participants to have a minimal understanding of such online workspace software.

3.5 Data collection and analysis

After the survey with 17 groups, we collected the recordings of their discussions and operations during the survey. To get the contextual factors and values involved in the usage of YouTube and justify the effectiveness of our method, we conducted a content analysis of the transcriptions of the recordings. The analysis is to answer the two research questions we proposed in the introduction. Therefore, the codes that we used to denote the transcriptions are from the following three perspectives:

- P1: Contextual factors influencing how much a participant likes a video
- P2: Values users may wish to receive from a recommender system.
- P3: Validation of our method: How the design principles helped us get the desired results.

Three rounds of coding were conducted for all the 20 groups, and each round is for a perspective above. Considering the subjectivity of coding human values, inter-coder reliability on value codes was calculated to improve the trustworthiness and rigor of our findings.

3.5.1 Initial codes

Generating initial codes is the first step for content analysis as it helps us break the transcriptions into manageable components. We began with reading all the transcriptions to get familiar with the contents. Based on a literature summarizing the most used contexts in HCI by Jumisko-Pyykkö and Vainio[27], we summarized the following initial codes for contexts, and a complete coding scheme of contexts can be found in Appendix B.1:

- Technical and informational Context: Technical aspects such as the device used.
- Cognitive Context: Habits, Self-Perception, Mood, etc.
- Physical Context: The environments that influence users
- Social Context: Influences from other people
- Temporal Context: Time-related factors and past experiences.

As for values, we went through J.Stray et al's synthesis about human values that are especially related to recommender systems[48]. These values are categorized into five themes: Usefulness, Well-being, Societal values, Public Discourse, and Legal and human rights. Under the five themes, We selected values that may occur in YouTube settings as a case study and coded the discussions involving the values in the transcriptions.

Theme	**Value**
Usefulness	Agency, Autonomy, Efficacy
Well-being	Well-being
	Connection
	Physical health
	Mental Health
	Community, Belonging
	Recognition, Acknowledgment
	Self-expression, authenticity
	Care, Compassion, Empathy
	Self-actualization, Personal Growth
	Inspiration, Awe
	Entertainment
Legal and Human Rights	Accessibility, Inclusiveness
Public Discourse	Accuracy (Factuality)
	Diversity
	Knowledge, Informativeness
Societal values	Progress
	Tradition, History

Table 3.1: Value themes

[48]

Chapter 4

Results

In this section, we will first demonstrate exemplary cases showing how context influences users' experiences and the values they want to attain in different scenarios while using recommender systems. Following the cases, We will enumerate the frequency of participant mentions for each context and corresponding value.

4.1 Influences of context

In this section, we present a series of cases that exemplify the influence of context when using recommender systems. These cases have been selected to provide a deep and novel understanding of the participant's experiences and perspectives.

4.1.1 The ongoing activity influences the selection of videos

While analyzing the transcriptions, many participants mentioned watching videos while doing other things as a form of companionship. We found participants who listened to some videos before sleeping to fall asleep more easily and also who watched videos while gaming or working to make them more engaging. From the examples, we can tell that video watching has the functionality of making the ongoing activity more enjoyable and enhancing the ongoing activity. We also found that users select different kinds of videos for companionship while doing different activities, fulfilling various emotional and cognitive needs.

Take the story of Cici as an example. She works for a consulting company and lives on her own. Whenever she returns home from work, she watches videos to relax. Even though she is usually tired after work, she still seeks educational content, as it provides her with knowledge and keeps her informed. During our survey, she shared an interesting experience: while preparing dinner, she watched a video about real estate, finding it both informative and enjoyable, which made the cooking process not boring. However, she immediately noted that she would not watch such a video while eating, as there is too much information and it would require too much cognitive effort. Instead, she would prefer more relaxing content, such as vlogs or travel videos, which provide comfort and ease without demanding active focus.

Cici's story illustrates that users often multitask while watching videos, using them as background companionship to enhance their experience of other activities. This behaviour reflects a broader trend where users select video content based on their current activity and cognitive state. Educational videos might fulfil a desire for self-improvement or productivity, while more relaxing content meets the need for emotional comfort and unwinding. Therefore, we can tell that videos not only serve as entertainment but also enrich daily routines by providing emotional support, mental stimulation, and even a sense of social connection.

4.1.2 Past experiences bring attraction to video preferences.

In addition to using videos to enhance their activities and find companionship, we also observed that participants are more easily drawn to content related to experiences they have had in the past. This familiarity often sparks a desire to learn more and deepen their understanding of these experiences. For example, one of our participants, Astro, a computer science student from Austria, primarily uses YouTube to acquire knowledge and stay informed about developments in his field. When asked to recommend a video to a friend, Astro selected one that explores how Central Park in New York City was developed. Astro mentioned that he had visited New York City several times but was unaware of the historical struggle involved in creating Central Park. The video, which detailed the challenges and decisions that shaped the park, intrigued him because it provided new insights into a place he had personally experienced.

This case illustrates how users are more likely to be attracted to content related to their experiences. Astro's prior visits to New York City made the video about Central Park particularly compelling, as it offered a deeper understanding of a familiar location. This finding suggests that recommender systems could enhance user engagement by suggesting content that builds on users' past experiences, making the recommendations more relevant and appealing.

4.1.3 Impact of cognitive context on video preferences

The two examples above are both external contexts, and this example shows how personal cognitive context influences users' experiences. One illustrative example is Mary, a participant who expressed a deep interest in computer games despite not being able to play them well herself. There are certain games she wishes to try, but due to their difficulty and her work. Therefore, she usually watches others playing the games she likes. She described this experience as a way of having escapism without committing to playing the game herself. This behaviour can be regarded as vicarious goal satisfaction. According to Kathleen C. McCulloch et al, vicarious goal satisfaction is "a phenomenon in which individuals experience "post-completion goal satiation" as a result of unwittingly taking on another person's goal pursuit and witnessing its completion."[36]. In our results, we found this phenomenon occurred several times.

Mary's experience highlights how vicarious goal satisfaction operates in the context of video consumption. By watching others playing games she loves, Mary is able to derive a sense of accomplishment and enjoyment. This allows her to engage with the content and fulfill her interest in gaming, even when she is not actively participating in the gameplay.

This example underscores the idea that watching others succeed in tasks or challenges can provide a similar sense of satisfaction as achieving the goal oneself. For Mary, and likely for many others, this vicarious satisfaction is a key motivator for consuming certain types of video content, particularly in areas where they feel less confident or capable. This finding suggests that recommender systems could benefit from identifying and promoting content that aligns with users' cognitive contexts, potentially enhancing user engagement by catering to this psychological need.

4.2 Effectiveness of the design principles

4.2.1 Overview

To prove the effectiveness of the approach we proposed, we prove the effectiveness of our design principles by how efficient they are in eliciting contexts and values related to recommender systems usage. Each task aligns with a specific design principle, meaning that the discussions generated during each task can be seen as stemming from that principle. Therefore, to prove that the design principles we proposed contributed to finding relevant context and values for evaluating recommender systems, we analyzed discussions to determine how the design principles facilitated participants' reflection on context and values.

We first counted the co-occurrences between context, values and the discussions. From Tables 4.1 and 4.2, we can see that principle 1, which provides the basis for participant reflection, performs well in eliciting values and context. However, we can find that principle 2, which prompts participants to merge experiences, did a good job of facilitating reflection on context while not doing well on values. With 105 discussions triggered by principle 2, there are 60 times of mentions of values during the discussions, but 106 times of mentions of context. Principle 3, which makes participants judge their experiences considering context, performed to our expectation on eliciting values. But, its performance in finding related context is not good, with only 20 mentions in 56 discussions. From the two tables, we got an overview of the effect of the three design principles we proposed. We can give a preliminary deduction that providing materials that people can refer to is the basis for them to reflect on context and values, and helps their thinking about insights of their experiences. Principle 2 and 3 have their speciality. Experiences merging(Principle 2) can help people do more thinking in the context of their experiences but is not very effective in finding values. As for letting participants make contextual judgements, we can make a prima facie case that it is useful on eliciting discussion about values. Since it's designed mainly for finding values, its poor performance of findings context is acceptable.

4.2.2 Exemplary Cases

In the last subsection, we got an overview of how the design principles perform for eliciting values and context. To demonstrate further details about how they exactly work, we selected several exemplary cases to provide insights.

4.2.3 Providing reflection basis facilitates communication

In the survey of Frost and Conrad, Frost was the more active one and he led the discussions, while Conrad was relatively introverted and confused when answering some questions. For example, when asked what influenced their YouTube watching habits, Conrad did not know how to start. Thanks to the visualizations, Frost helped Conrad summarize that he woke up late and presumed that he didn't like serious videos after waking up. This triggered Conrad's thinking and gave him a foundation to begin with and eventually made good reflections. Based on Frost's summary, Conrad continued that his occupation is quite serious, so when surfing YouTube, he prefers light and entertaining videos. He even extended another experience from this that he once made a mistake at work and he was very sad. Therefore, he watched a prank video about people hiding behind bushes scaring other people crossing the streets for two hours and it really lightened his mood.

From this example, we see that providing a reflection basis could improve the data quality when interviewing introverted participants who are not talkative or confused about where to start. It provides possibilities for leader participants to help their teammates to find an answer.

4.2.4 Merging experiences supports cooperative reflection

In the survey of Lily and Drew, they shared the same interests, so they had active discussions. They even exchanged Instagram accounts after the survey. So, their discussions were chosen to demonstrate how experience merging facilitates reflection. When they were asked to recommend a video to each other, Drew recommended an MV of a song called "The Coconut Song", and she recommended it because this video can bring happiness. After hearing this, Lily was so surprised that she said: "I'm gonna scream out." She immediately responded that she also loves this song and she always plays this song whenever she drives a long distance. Lily even reflected on the values she got by listening to this song while our question did not even ask. She said that listening to this song clears the tiredness and brings happiness to the boring journey.

From this example, when participants find similarities with each other, they can connect their own experiences with others, which helps elicit the insights behind the behaviours.

4.2.5 Making contextual judgments inspires deep thinking of the scenario and values

In this example, we will introduce the story of Judy. She works as an assistant at a law firm in Spain. She said that most of the time, she feels exhausted after work. Therefore, she usually picks videos that cheer her up and empty her mind. When asked to judge whether she was satisfied with a video she watched. She said that she was disappointed because she played the video for her naughty younger brother such that she did not have to deal with him. However, she was hoping that even if it is for kids, it could be enjoyable for adults to some extent, at least not annoying. It happened after she arrived home, which made her desire to relax stronger than normal moments. This made it even more frustrating that the video only helped her manage the kids but didn't cheer her up.

By prompting her to judge the video considering the context, we got an in-depth understanding of the scenarios when she watched the video and the values she wished to receive in the context. We know the context of YouTube watching that users play videos for children to calm them down. In this context, users may seek more than just the relief of not having to manage their children. They may desire content that offers personal benefits, such as improving mental health, providing inspiration, or simply offering entertainment.

	Public Discourse (88)	**Usefulness (39)**	**Well-being (146)**	**Total**
Principle 1 (182)	34	20	65	119
Principle 2 (105)	15	7	38	60
Principle 3 (56)	22	12	34	68
Total	90	41	151	282

Table 4.1: Co-occurences between values and design principles.

	Application Context (16)	**Cognitive Context (55)**	**Physical Context (33)**	**Social Context (50)**	**Temporal Context (103)**	**Total**
Principle 1 (182)	9	24	16	21	75	145
Principle 2 (105)	7	21	16	23	39	106
Principle 3 (56)	1	8	1	4	6	20
Total	17	53	33	48	120	271

Table 4.2: Co-occurences between context and design principles.

4.3 Contexts

4.3.1 Contexts occurrences

To have an overview on what are the most influential contextual factors, we first counted the total number of occurrences of each context. From Figure 4.1, we can see that time, working/study status, and eating are the top 3 factors that were mentioned the most among the 17 groups of participants. Also, we can see that time and working/study status are way ahead of others. Time was mentioned 78 times, and work/study status was mentioned 56 times, and they constituted 36.81% of total occurrences(364 times)

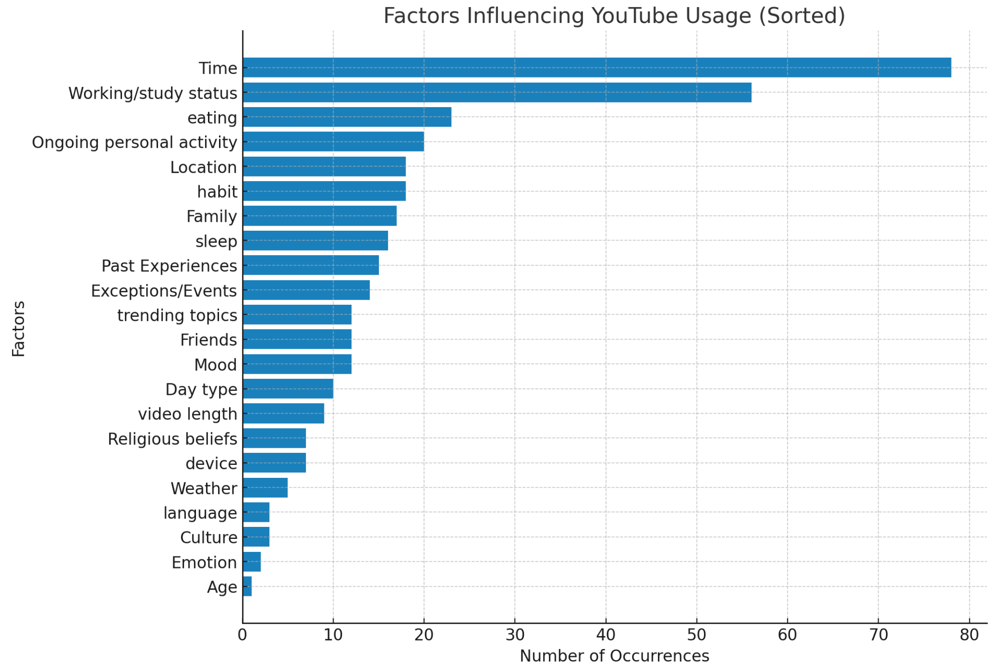


Figure 4.1: Total occurrences of contextual factors

4.3.2 Sample quotes for each context

To gain a clearer understanding of the influence of each factor, we also tracked the number of individuals who mentioned each factor and provided a representative quote for each context. The sample quotes should give an overview of how each context influences users watching experience. As we did our survey with 17 groups in total, the total number of participants is 34. As shown in Table 4.1, both time and work/study status were mentioned by half of the participants. Furthermore, location, eating, past experiences, habits, events, and sleep were among the other factors mentioned by more than 10 participants.

Context	**Number of Participants(N=32)**	**Sample Quote**
Time	31	“Cause usually in the evenings like I’m usually watching, maybe interviews or maybe some podcasts or maybe some guidance on how to do something like when I’m in bed. So that’s when I usually watch.”

Context	Number of Participants	Sample Quote
Work/study status	24	"Yeah, I mean I have like a I've got quite an intense job. I work at a law firm. So I think a lot of the time I'm watching YouTube to, like, relax, I guess, and not. So I'm definitely like escapism, relaxation because I don't want to think about work. So I like turn."
Location	15	"I was just traveling on the way back home, and I wanted to listen to something and chill out."
Eating	14	"OK, it was my lunch time at work, so I spent it watching that video. Isolating myself from my colleagues."
Past Experiences	12	"Why they take the spot that they picked for the park, how they got the land and sort of how New York City looked like hundreds of years ago without any sort of like public green space, it's really it's really interesting. And I've been to New York City a few times and I had no idea they fought so hard for that."
Habit	11	"I wake up slowly in general, so I need a couple of hours of, you know, watching videos and listening to music."
Events	11	"So this is about, yeah, it's car maintenance and I bought a car recently and I want to learn how to do mechanical jobs to save money."
Sleep	11	"As you know, to let you go asleep to make you fall asleep easier quite often, I wouldn't even physically watch them. I would just listen to them and then, you know, I go and I fall asleep eventually. And those are all different type of videos. Sometimes educational sometimes."
Trending topics	9	"Because I'm in sort of computer science, I sort of have to watch a lot of videos to sort of see what's new. Like I watch I learn about like AI or or blockchain like crypto."
Friends	9	"I would watch it, but only when I'm doing it with other people as well, because I don't usually watch long videos about comedy or something by myself."
Video Length	8	"Because they are short and they are, you know, constant dopamine hits."

Context	Number of Participants	Sample Quote
Day type	7	"I usually have a certain schedule mostly on Sundays, because I receive newsletters mostly on Sundays, so I just go through some of the things that other developers have done during the week, so I usually do the catch-up on Sunday or through. So my my learning videos are usually."
Mood	7	"It was people hidden behind bushes scaring other people crossing the streets. So I was at a very low point at that moment. I was feeling kind of, you know. Just Mystic and sad. So I wanted something funny to lighten my mood."
Device	6	"I went out to the shopping site I used my phone to watch videos and listening to music from YouTube. And then when I came back I switched to my laptop."
Family	6	"I think for me right before work I put on like this kids YouTube channel because I have a I have a baby brother and he was really annoying me. So I just wanted to make him watch the video and not have to deal with him so obviously."
Weather	5	"I'm trying to look into C# development etc, but currently it's summer as you've said, yeah, watching more like comedy videos."
Religious beliefs	4	"I already said I'm a believer, so I also like just watching some of the sermons. And I like to be there and actually just listen, suck them in and understand, like a sense of community as well."
Culture	3	"I like. I really like music. What type of music do you listen to? Everything, everything I think except something traditional like some, you know, Serbia, Macedonia, Russia."
Video Type	2	"Yeah. It's like when you watch it, you can also learn something new. I think while I was preparing my dinner. If I'm eating my dinner, maybe I don't want to watch that, because it's too information-heavy."

Table 4.3: Contexts and Sample Quotes from Participants

To summarize the results we found for the contexts, this study identified several key contextual factors that shape YouTube usage, ranging from time of day to work commitments and social influences. We found that physical and temporal contexts and the two most influential contexts when users decide what videos to watch. As for temporal context, the most reported factor is time, particularly during evening routines and specific times of the day when participants are more likely to consume content. This suggests that YouTube viewing habits are often tied to specific times of day, and users tend to consume different kinds of content at different times. Physical context such as ongoing activities makes big differences. One significant example is this quote: ”Yeah. I like this video because, like when you watch this video, you can also learn something new. I think while I watched it when I was preparing my dinner. If I’m eating my dinner, maybe I don’t want to watch that, because there is a lot of information.” From this sample quote, we can tell that the participant liked the video but she would not watch it in some situations where she only wanted some relaxing content. As for the contexts that were less frequently mentioned, although they were not mentioned as much as the other ones, we found some unexpected points. Using day type as an example, one participant gets his newsletters on Sunday so Sunday is the time he uses YouTube to catch up on the latest technology developments by watching videos. In conclusion, contexts have a significant influence on the content that users wish to consume.

4.4 Values

4.4.1 Value occurrences

Similar to the analysis of contextual factors, we first counted the number of occurrences of each value. From Figure 4.2, we can see that the most frequently cited factor is ”Knowledge and Informativeness,” with 67 mentions. This suggests that participants highly value content that enhances their understanding and provides them with useful, informative insights. Followed by personal growth, which takes the second place with 44 mentions. Entertainment, companionship, mental health, and agency have similar mentions, following informativeness and personal growth. In summary, the results indicate a strong preference for content that is informative, promotes personal growth and provides entertainment. The importance of mental health, companionship, and agency further underscores the multifaceted role of media in supporting well-being and fostering meaningful public discourse. The emphasis on diversity and accuracy also highlights the audience’s demand for inclusive and trustworthy content, which are essential for maintaining credibility and engagement in public conversations.

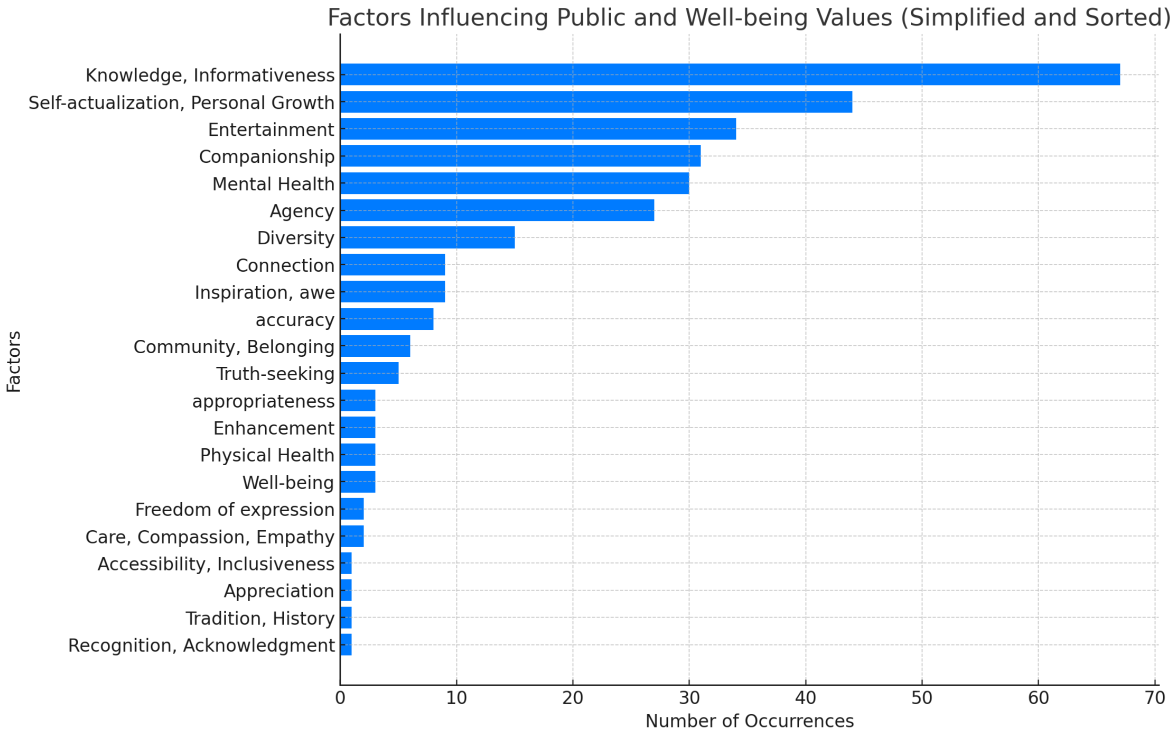


Figure 4.2: Total occurrences of values

According to J. Stray et al in their synthesis about human values in recommender systems[48], these values can be grouped into several themes: Usefulness, Well-being, Legal and human rights, public discourse, and societal values. After grouping the values into several themes, we found that the values most users want to receive from YouTube are well-being, public discourse, and usefulness. As for legal and human rights, and societal values, there is only one mention for each.

4.4.2 Sample quotes for each value

To further understand how participants require these values, we count the number of participants who reported each value and show sample quotes for each value. The data in Table 4.2 reveals that YouTube fulfills users' needs for self-improvement, education, entertainment, goal-achieving, and companionship. The five values are the ones that are mentioned by half of the participants.

Values	Number of participants mentioned(N=32)	Sample quote

Self-actualization, personal growth	26	“Because I’m in sort of computer science, I sort of have to watch a lot of videos to sort of see what’s new. Like I watch I learn about like AI or or blockchain like crypto.”
Knowledge, Informativeness	25	“I sort of watch, like educational stuff about like technology and documentaries and news. I try not to watch like silly things. I use it to just sort of learn.”
Entertainment	21	“But currently it’s it’s summer as you’ve said, yeah, watching more like comedy videos.”
Agency	17	“I can say the for the flies. I’ve learned how to kill them. So how to how to get the job done.”
Companionship	17	“I watch YouTube videos, particularly as company my meals.”
Mental Health	14	“Yeah, I mean I I have like a I’ve got quite an intense job. I work at a law firm. So I think a lot of the time I’m watching YouTube to, like, relax, I guess, and not. So I’m definitely like escapism, relaxation because I don’t want to think about work. So I like turn.”
Diversity	10	“I like learning from different people. You give me the chance to learn from like different people, different experts. So I prefer I prefer learning from YouTube than buying courses because when I buy a course I buy it from just one person.”
Inspiration	8	“And then I like photography a lot and then graphic designing. So sometimes I watch some tutorials on them to expand my creativity.”
Connection	7	“I already said I’m a believer, so I also like just watching some of the sermons. And I like to be there and actually just listen, suck them in and understand.”
Accuracy	6	“so it was a video that I watched that was explaining a new habit that was developed in some countries where people have started to exhibit violence against the domestic. So it was really surprising to me to view it and then towards the end, the video announced the fact that this kind of information only talks about a small percentage of people in these countries. So it was really frustrating.”

Community, Belonging	5	"They'll call Ahmadiyya Muslim community and the argument is based on Quran and Bible because they believe that the second coming of Jesus Christ, they believe that the those guys have come back in the second time and the 2nd. Their father founded their community. It let me learn about the community."
Truth-seeking	4	"I was expecting to get, you know, some precise numbers. And, you know, some analysis of the subject in general. And this is not by, you know, talking about percentages. When I started watching the video, it was like the majority of people do this. Like this common practice. And then we were talking about 2% of the population. So the values that I didn't get is accuracy I suppose, and truthfulness which was not there."
Physical Health	2	"I watch mainly for exercise and stretching and things like that."
Freedom of expression	2	"So I find that YouTube is very neutral and you can learn about a lot of things and not political way. It's like almost nobody really fights. About it, and if you read like the comments, then people are usually very respectful."
Care, Compassion, Empathy	2	"It gives you my spiritual strength. And it helped me to be strong in my lives. So that's all."
Appropriateness	2	"I had like a kid with me at home or I needed to put something on YouTube that that was like, safe and not inappropriate then."
Accessibility, Inclusiveness	1	"So I find that YouTube is very neutral and you can learn about a lot of things and not political way. It's like almost nobody really fights about it, and if you read like the comments, then people are usually very respectful and I find it a lot nicer to do that on YouTube like reading like an article or watching like news on TV. So that's sort of why I choose YouTube for that."
Tradition, history	1	"Also we spoke about being growing up Catholic and I have always been fascinated about the history in Catholic."

Recognition	1	“The video that made it most satisfying for me was fact to say, drove. It made me want to play longer. And again, it just it brings me joy and I think that just helps me feel a lot better about what I’m what I’m playing.”

Table 4.4: Value sample quotes

4.4.3 Inter-coder reliability

As mentioned in the methods section, coding values is a subjective process; therefore, inter-coder reliability was calculated to ensure consistency in the coding. To evaluate this, two coders independently coded five documents, and the inter-coder reliability was calculated based on these documents. The Holsti index was chosen to calculate inter-coder reliability because it is particularly well-suited for measuring the agreement between two coders across multiple categories. This index provides a straightforward and effective method for assessing the consistency of coding decisions. From Figure 4.3, we can see that the Holsti index of the two coders’ results on values is 35.4%. Considering the coding style between the two codes is different: One coded the exact sentence that demonstrated the value while the other coded the complete discussions related to the sentence, this result reached our expectations.

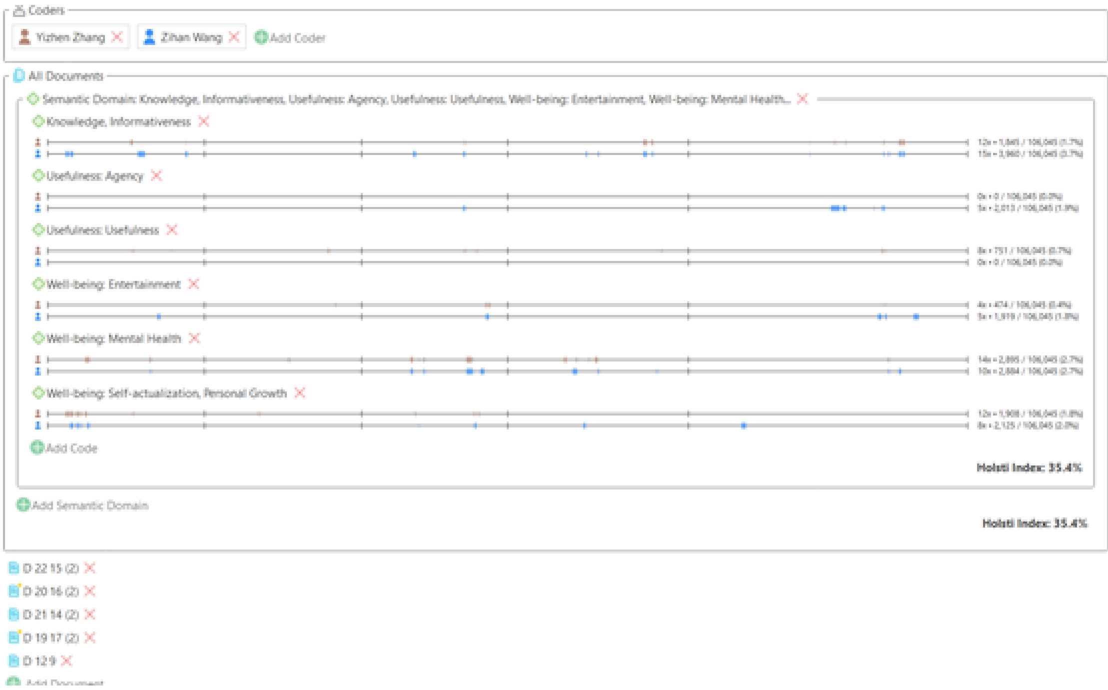


Figure 4.3: Inter-coder reliability test: Holsti Index

4.5 Method Usability

Since we proposed a crowdsourcing method, the usability of our method is important to be mentioned. As we mentioned in the method section(Section 3.4), the platform did not work as well as we expected on crowdsourcing platforms in the very beginning. After the measure we mentioned, the results significantly improved and we counted the number of participants who left at each stage of the tasks in the formal test and listed it below.

- **Total Number of Participants (N)**: 26
- **Participants left in the Waiting Room**: 6
- **Paired successfully but didn't Join the Miro Board**: 5
- **Participants left after entering the Miro board**: 7

We can tell from the list that few participants left because of the long waiting time. Although there were still some participants who quit after being paired, our re-assign mechanism ensured that the left participant got a new teammate in no more than five minutes. This avoided the situation that one participant had to go back to the waiting room if his teammate left. All in all, our method's usability reached our expectations on crowdsourcing platforms after making adjustments based on what we learned from pilot tests.

Chapter 5

Discussion, Conclusions, and Future Work.

This chapter gives an overview of the project's contributions. After this overview, we will reflect on the results and draw some conclusions. Finally, some ideas for future work will be discussed.

5.1 Discussion

The results of our studies provided valuable insights into the influence of context on users' experiences with recommender systems. At the same time, these findings demonstrate that collaborative reflection plays a critical role in effectively evaluating user experiences by offering ample support for thoughtful engagement. This section will further explore the results to answer the two primary research questions.

To answer the first question, we analyzed how different contexts influenced users' choices of video types and the values they wanted to get. The results revealed that context related to users' daily routines had the greatest influence on their choice of YouTube videos such as work or study environments, eating, and sleep patterns. The values in these contexts mainly are informativeness, self-actualization, entertainment and companionship. When users are doing different activities, the purposes of a YouTube video being played vary. This is embodied in Cici's case and others involving watching YouTube while doing another activity. At the same time, these daily routines highly correlate with other contexts such as location, time, socials, etc. One participant reported that he usually spends one hour in the morning enjoying coffee and watching YouTube to wake himself up and get ready for work. Therefore, recommending videos that boost his energy when he wakes up would enhance his experience. This experience involves the context of daily routine, time, and working conditions. It co-occurs with the values of inspiration and entertainment. From this, we concluded the first point to consider when evaluating the recommendations of recommender systems: We do not judge only on whether the systems give content that users are interested in. We should examine whether the contents are the most appropriate ones in the scenario by checking whether the recommended items at different times can fulfil the user's values. It should align with users' daily routines and give the most appropriate recommendations at each period instead of giving similar ones across the day.

In the previous paragraph, we talked about how we should consider daily routine-related context while evaluating recommender systems. However, we also found contexts that do not occur regularly. For example, trending topics, events, and mood. Compared to the daily-routine contexts, they are more unpredictable. Therefore, we can greatly improve user experiences if we can capture this unusual context and give appropriate recommendations. On the other hand, giving unsatisfying recommendations in these scenarios brings more frustration to users. One of our participants reported a case where he opened a video to cheer him up when he was upset. The video began well but gradually became off-topic, even with promotions in the video, which made him even more stressed. Although these contexts are relatively random, we still found traces from participants' discussions. The man was upset because of his tiring work, another participant got interested in cliff diving because Red Bull was organizing an event in his village, etc. We can notice that although these contexts are random, their causes can be connected to a stable context. The man's frustration stemmed from his work, while the Red Bull event is linked to the participant's location, which serves as a stable key context. Therefore, we gave the second conclusion: When evaluating user experiences on recommender systems, we should examine whether the systems can capture unusual context.

The last part of the context are users' cognitive context such as culture, religious beliefs, etc. These contexts are stable and influence a part of users who are heavily influenced. Therefore, deploying measures of knowing users' cognition and considering it when giving recommendations is the last point that we concluded to evaluate recommender systems.

In this paragraph, we will answer the second research question on how collaborative reflection helps find insights into user experiences. From the design principles-value and design principles-context co-occurrence results, we verified the effectiveness of collaborative reflection. We concluded that by giving enough support on materials to reflect on, experiences merging, and contextual judgment, participants can reflect on their past experiences smoothly and think of values logically. The design principles we proposed made the reflection process natural and aligned with human thinking. The first design principle ensures that participants are well-acquainted with their own data, forming the basis for principles 2 and 3, where experiences are integrated and judged concerning context. The collaborative setting makes the process of examining their data not just an input process, but a process in which users get their data as input and output their own reflections. The existence of their teammates makes the reflection results brought out in natural discussions. In conclusion, collaborative reflection prompts participants to reflect on their experiences by talking to other people instead of answering boring questions one by one. It makes the process more interactive and lively and turns out to be effective in finding insights into people's experiences.

However, this approach has limitations. The first one is its requirements for participants, especially in an online environment. Our study did not go as smoothly as we expected in the beginning because many participants were not willing to talk to a stranger in an online video chat. This not only led to frequent withdrawals by participants but also the depth of communication. We addressed the issue by configuring the filter to accept only participants who are willing to participate in a video interview and share the same interests. The second limitation of our study is the lack of diversity in our participant groups. Most of our participants aged from twenty to thirty and most of them know about operating computers well and can finish the interactive tasks without too much confusion. However, the approach did not work well with participants who could not use computers fluently. Therefore, the effectiveness of the approach on older population groups can not be verified.

5.2 Conclusions

Recommender systems are the foundation of many internet companies and whether or not to be able to give satisfying recommendations determines the product quality. Context influences have been recognised for recommender systems[3] and other areas[45]. Therefore, understanding the details of how context influences user experiences is the first step to evaluating a recommender system and making further improvements accordingly.

The contributions of this project are two parts: It fills the knowledge gap of there is no effective method for evaluating how a recommender system deals with user context with a user study approach using collaborative reflection. Along with the design principles we proposed, the approach effectively finds insights using collaborative reflection. It prompts people to reflect on past experiences and eventually inspire deep thinking on values. We concluded that by providing enough support on communication and experiences merging, it performs well in evaluating recommender systems experiences. The second is that we found how context influences users' experiences on recommender systems and the values they wish to be fulfilled under different contexts. We summarized the context into three categories: Daily routine, unusual events, and personal cognitive context. For each category, we give recommendations on how to evaluate a recommender system's measure on them. These findings not only provide actionable guidelines for developers to evaluate and improve their systems but also highlight the potential for more personalized and context-aware recommendations.

Our approach effectively uncovered valuable insights, and the possible ways to evaluate how the context is involved in giving recommendations. Future studies could benefit from a larger and more diverse participant pool. Also, research on how to include these contexts comprehensively in the algorithms and measures to improve user satisfaction considering these contexts is valuable. Overall, this research contributes an approach to finding insights into user experiences. It also gives guidance on how to evaluate recommender systems' measures in context. It contributes to advancing context-aware recommender systems and lays the groundwork for more nuanced and adaptive recommendation strategies in the future.

Bibliography

[1] Recommender system — Wikipedia, the free encyclopedia, 2024. URL `https://en.wikipedia.org/wiki/Recommender_system`.

[2] Stav Atir, Xuan Zhao, and Margaret Echelbarger. Talking to strangers: Intention, competence, and opportunity. *Current Opinion in Psychology*, 51:101588, 2023.

[3] Linas Baltrunas, Bernd Ludwig, and Francesco Ricci. Context relevance assessment for recommender systems. In *Proceedings of the 16th international conference on Intelligent user interfaces*, pages 287–290, 2011.

[4] Tara S Behrend, David J Sharek, Adam W Meade, and Eric N Wiebe. The viability of crowdsourcing for survey research. *Behavior research methods*, 43:800–813, 2011.

[5] Frank Bentley, Konrad Tollmar, Peter Stephenson, Laura Levy, Brian Jones, Scott Robertson, Ed Price, Richard Catrambone, and Jeff Wilson. Health mashups: Presenting statistical patterns between wellbeing data and context in natural language to promote behavior change. *ACM Trans. Comput.-Hum. Interact.*, 20(5), nov 2013. ISSN 1073-0516. doi: 10.1145/2503823. URL `https://doi.org/10.1145/2503823`.

[6] Andrew BL Berry, Catherine Y Lim, Calvin A Liang, Andrea L Hartzler, Tad Hirsch, Dawn M Ferguson, Zoë A Bermet, and James D Ralston. Supporting collaborative reflection on personal values and health. *Proceedings of the ACM on human-computer interaction*, 5(CSCW2):1–39, 2021.

[7] Ram D. Gopal Rajkumar Venkatesan Bhavik Pathak, Robert Garfinkel and Fang Yin. Empirical analysis of the impact of recommender systems on sales. *Journal of Management Information Systems*, 27(2):159–188, 2010. doi: 10.2753/MIS0742-1222270205. URL `https://doi.org/10.2753/MIS0742-1222270205`.

[8] Alan Borning and Michael Muller. Next steps for value sensitive design. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1125–1134, 2012.

[9] Jorge Carvalho, Arnaldo Santos, and Hugo Paredes. Data quality improvement in crowdsourcing systems by enabling a positive personal user experience. In *2019 IEEE 23rd International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 255–260. IEEE, 2019.

[10] Janghee Cho, Tian Xu, Abigail Zimmermann-Niefield, and Stephen Voida. Reflection in theory and reflection in practice: An exploration of the gaps in reflection support among personal informatics apps. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. doi: 10.1145/3491102.3501991. URL `https://doi.org/10.1145/3491102.3501991`.

[11] Eun Kyoung Choe, Bongshin Lee, Haining Zhu, Nathalie Henry Riche, and Dominikus Baur. Understanding self-reflection: how people reflect on personal data through visual data exploration. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare*, PervasiveHealth '17, page 173–182, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450363631. doi: 10.1145/3154862.3154881. URL `https://doi.org/10.1145/3154862.3154881`.

[12] Naomi F Dale, Raechel Johns, and Michael James Walsh. "that's not true!" paired interviews as a method for contemporaneous moderation of self-reporting on a shared service experience. *Journal of Hospitality and Tourism Management*, 49:580–591, 2021.

[13] Marilyn Wood Daudelin. Learning from experience through reflection. *Organizational dynamics*, 24(3):36–48, 1996.

[14] Esther de Groot, Maaike D Endedijk, A Debbie C Jaarsma, P Robert-Jan Simons, and Peter van Beukelen. Critically reflective dialogues in learning communities of professionals. *Studies in Continuing Education*, 36(1):15–37, 2014.

[15] Pieter MA Desmet and Sabine Roeser. Emotions in design for values. *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains*, 203:219, 2015.

[16] Jiaqi Fan, Weimin Pan, and Lisi Jiang. An improved collaborative filtering algorithm combining content-based algorithm and user activity. In *2014 International Conference on Big Data and Smart Computing (BIGCOMP)*, pages 88–91, 2014. doi: 10.1109/BIGCOMP.2014.6741413.

[17] Batya Friedman, Peter Kahn, and Alan Borning. Value sensitive design: Theory and methods. *University of Washington technical report*, 2(8):1–8, 2002.

[18] Iason Gabriel. Artificial intelligence, values, and alignment. *Minds and machines*, 30 (3):411–437, 2020.

[19] G Geetha, M Safa, C Fancy, and D Saranya. A hybrid approach using collaborative filtering and content based filtering for recommender system. *Journal of Physics: Conference Series*, 1000(1):012101, apr 2018. doi: 10.1088/1742-6596/1000/1/012101. URL `https://dx.doi.org/10.1088/1742-6596/1000/1/012101`.

[20] Google. *YouTube Data API v3*, n.d. URL `https://developers.google.com/youtube/v3/docs`. https://developers.google.com/youtube/v3/docs.

[21] Lisa Graham, Anthony Tang, and Carman Neustaedter. Help me help you: Shared reflection for personal data. In *Proceedings of the 2016 ACM International Conference on Supporting Group Work*, GROUP '16, page 99–109, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342766. doi: 10.1145/2957276.2957293. URL `https://doi.org/10.1145/2957276.2957293`.

[22] Lisa Graham, Anthony Tang, and Carman Neustaedter. Help me help you: Shared reflection for personal data. In *Proceedings of the 2016 ACM International Conference on Supporting Group Work*, pages 99–109, 2016.

[23] Mokter Hossain and Ilkka Kauranen. Crowdsourcing: a comprehensive literature review. *Strategic Outsourcing: An International Journal*, 8(1):2–22, 2015.

[24] Yifan Hu, Yehuda Koren, and Chris Volinsky. Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE international conference on data mining*, pages 263–272. Ieee, 2008.

[25] Kosetsu Ikeda, Atsuyuki Morishima, Habibur Rahman, Senjuti Basu Roy, Saravanan Thirumuruganathan, Sihem Amer-Yahia, and Gautam Das. Collaborative crowdsourcing with crowd4u. *Proceedings of the VLDB Endowment*, 9(13):1497–1500, 2016.

[26] Gawesh Jawaheer, Peter Weller, and Patty Kostkova. Modeling user preferences in recommender systems: A classification framework for explicit and implicit user feedback. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 4(2):1–26, 2014.

[27] Satu Jumisko-Pyykkö and Teija Vainio. Framing the context of use for mobile hci. *International journal of mobile human computer interaction (IJMHCI)*, 2(4):1–28, 2010.

[28] Aniket Kittur, Ed H Chi, and Bongwon Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 453–456, 2008.

[29] Bart P Knijnenburg, Martijn C Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. Explaining the user experience of recommender systems. *User modeling and user-adapted interaction*, 22:441–504, 2012.

[30] Anand Kulkarni, Matthew Can, and Björn Hartmann. Collaboratively crowdsourcing workflows with turkomatic. In *Proceedings of the acm 2012 conference on computer supported cooperative work*, pages 1003–1012, 2012.

[31] Albrecht Kurze, Andreas Bischof, Sören Totzauer, Michael Storz, Maximilian Eibl, Margot Brereton, and Arne Berger. Guess the data: Data work to understand how people make sense of and use simple sensor data from homes. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–12, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376273. URL `https://doi.org/10.1145/3313831.3376273`.

[32] Ian Li, Anind K. Dey, and Jodi Forlizzi. Understanding my data, myself: supporting self-reflection with ubicomp technologies. In *Proceedings of the 13th International Conference on Ubiquitous Computing*, UbiComp '11, page 405–414, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450306300. doi: 10.1145/2030112.2030166. URL `https://doi.org/10.1145/2030112.2030166`.

[33] Zhan Liu, Shaban Shabani, Nicole Glassey Balet, Maria Sokhn, and Fabian Cretton. How to motivate participation and improve quality of crowdsourcing when building accessibility maps. In *2018 15th IEEE Annual Consumer Communications & Networking Conference (CCNC)*, pages 1–6. IEEE, 2018.

[34] R Logesh and V Subramaniyaswamy. Exploring hybrid recommender systems for personalized travel applications. In *Cognitive Informatics and Soft Computing: Proceeding of CISC 2017*, pages 535–544. Springer, 2019.

[35] Narges Mahyar and Melanie Tory. Supporting communication and coordination in collaborative sensemaking. *IEEE transactions on visualization and computer graphics*, 20(12):1633–1642, 2014.

[36] Kathleen C McCulloch, Gráinne M Fitzsimons, Sook Ning Chua, and Dolores Albarracín. Vicarious goal satiation. *Journal of Experimental Social Psychology*, 47(3): 685–688, 2011.

[37] Miro. *Miro Web SDK Reference*, 2024. URL `https://developers.miro.com/docs/sdk-reference`. Accessed: 2024-07-12.

[38] Marion K Poetz and Martin Schreier. The value of crowdsourcing: can users really compete with professionals in generating new product ideas? *Journal of product innovation management*, 29(2):245–256, 2012.

[39] Michael Prilla and Bettina Renner. Supporting collaborative reflection at work: A comparative case analysis. In *Proceedings of the 2014 ACM International Conference on Supporting Group Work*, pages 182–193, 2014.

[40] Aare Puussaar, Adrian K Clear, and Peter Wright. Enhancing personal informatics through social sensemaking. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 6936–6942, 2017.

[41] Elizabeth C Ravlin and Bruce M Meglino. Effect of values on perception and decision making: A study of alternative work values measures. *Journal of Applied psychology*, 72(4):666, 1987.

[42] Daniel M. Russell, Mark J. Stefik, Peter Pirolli, and Stuart K. Card. The cost structure of sensemaking. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*, CHI '93, page 269–276, New York, NY, USA, 1993. Association for Computing Machinery. ISBN 0897915755. doi: 10.1145/169059.169209. URL `https://doi.org/10.1145/169059.169209`.

[43] J Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. Collaborative filtering recommender systems. In *The adaptive web: methods and strategies of web personalization*, pages 291–324. Springer, 2007.

[44] Shalom H Schwartz. Basic human values: An overview. 2006.

[45] Neil Gareth Shepherd and John Maynard Rudd. The influence of context on the strategic decision-making process: A review of the literature. *International journal of management reviews*, 16(3):340–364, 2014.

[46] Steven M Smith, Arthur Glenberg, and Robert A Bjork. Environmental context and human memory. *Memory & Cognition*, 6(4):342–353, 1978.

[47] Jonathan Stray, Ivan Vendrov, Jeremy Nixon, Steven Adler, and Dylan Hadfield-Menell. What are you optimizing for? aligning recommender systems with human values. *arXiv preprint arXiv:2107.10939*, 2021.

[48] Jonathan Stray, Alon Halevy, Parisa Assar, Dylan Hadfield-Menell, Craig Boutilier, Amar Ashar, Chloe Bakalar, Lex Beattie, Michael Ekstrand, Claire Leibowicz, et al. Building human values into recommender systems: An interdisciplinary synthesis. *ACM Transactions on Recommender Systems*, 2(3):1–57, 2024.

[49] G Suganeshwari and SP Syed Ibrahim. A survey on collaborative filtering based recommendation system. In *Proceedings of the 3rd international symposium on big data and cloud computing challenges (ISBCC–16')*, pages 503–518. Springer, 2016.

[50] Alice Thudt, Uta Hinrichs, Samuel Huron, and Sheelagh Carpendale. Self-reflection and personal physicalization construction. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–13, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356206. doi: 10.1145/3173574.3173728. URL `https://doi.org/10.1145/3173574.3173728`.

[51] Alice Thudt, Uta Hinrichs, Samuel Huron, and Sheelagh Carpendale. Self-reflection and personal physicalization construction. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–13, 2018.

[52] Karthikeyan Umapathy. Requirements to support collaborative sensemaking. In *CSCW CIS Workshop*, volume 10, 2010.

[53] Mija Van Der Wege, Jessi Jacobsen, Nicole Magats, Carl Bou Mansour, and Jin Hong Park. Familiarity breeds overconfidence: Group membership and shared experience in the closeness-communication bias. *Journal of Experimental Social Psychology*, 94: 104097, 2021.

[54] Renjie Zhou, Samamon Khemmarat, and Lixin Gao. The impact of youtube recommendation system on video views. IMC '10, page 404–410, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781450304832. doi: 10.1145/1879141.1879193. URL `https://doi.org/10.1145/1879141.1879193`.

Appendix A

Glossary

A.1 Questions for task 1

1. Reflecting on Your Guess About Watching Behavior
   - What is your guess about your watching behavior, and why?
   - Did you find any interesting differences or similarities between your initial guess and the actual data? Why do you find them interesting?
2. Reflect on Different Data Patterns and Personal Contexts
   - Do you find any interesting similarities or differences between your data and your partner's data?
   - What is your current life situation (e.g., work, study, personal life), and what personal goals (e.g., relaxation, learning, escapism) drive your watching behavior?
   - How does your video-watching behavior represented by the data align with your personal values? Why?

A.2 Questions for task 2

1. Point out one video that satisfied you the most.
   - In what context did you watch this video, and why did you watch it??
   - What element of the video makes it the most satisfying for you?
   - Why are you satisfied by this particular element? How does it relate to your personal values?
2. Point out the video that satisfied you the least.
   - In what context did you watch this video, and why did you watch it?

- What elements of the video made you feel less satisfied?
- How do you value your watching behavior in such context, and why?

3. Pick one video that would interest your partner, and discuss the questions below.

- Please explain the reason for your recommendation.
- Please evaluate if you would like to watch the video recommended by your partner, and why.

A.3 Tasks

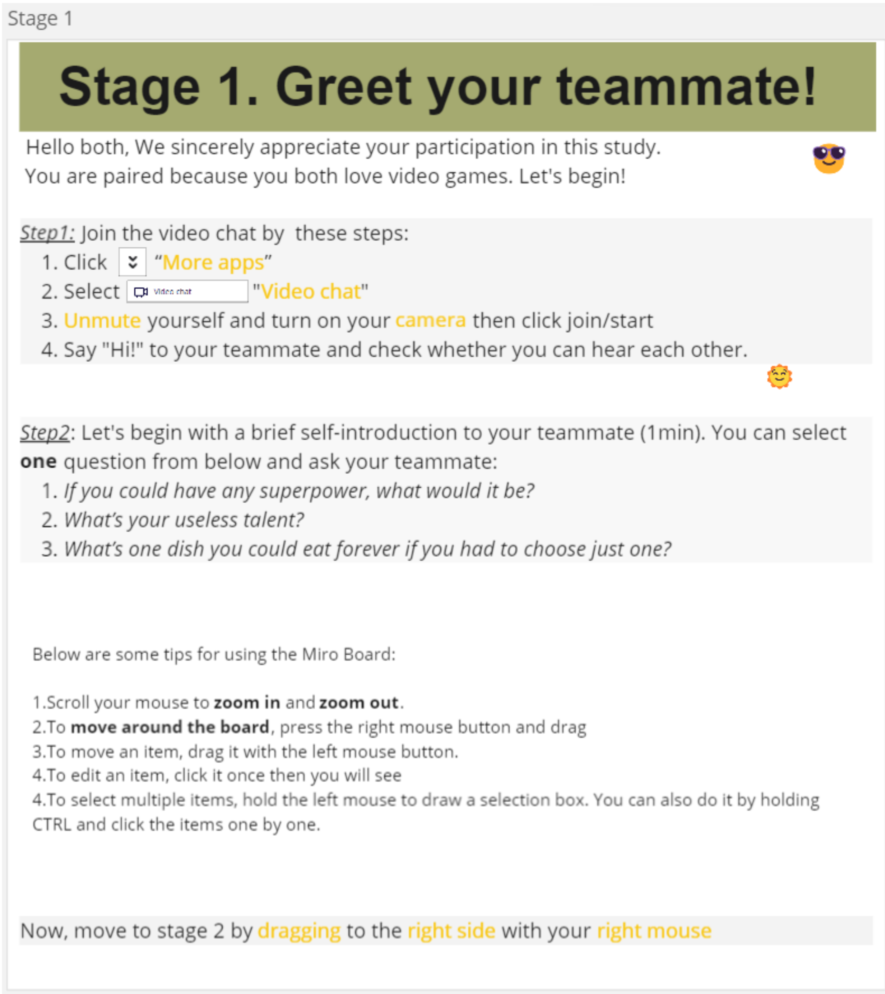


Figure A.1: Stage 1: Ice breaking

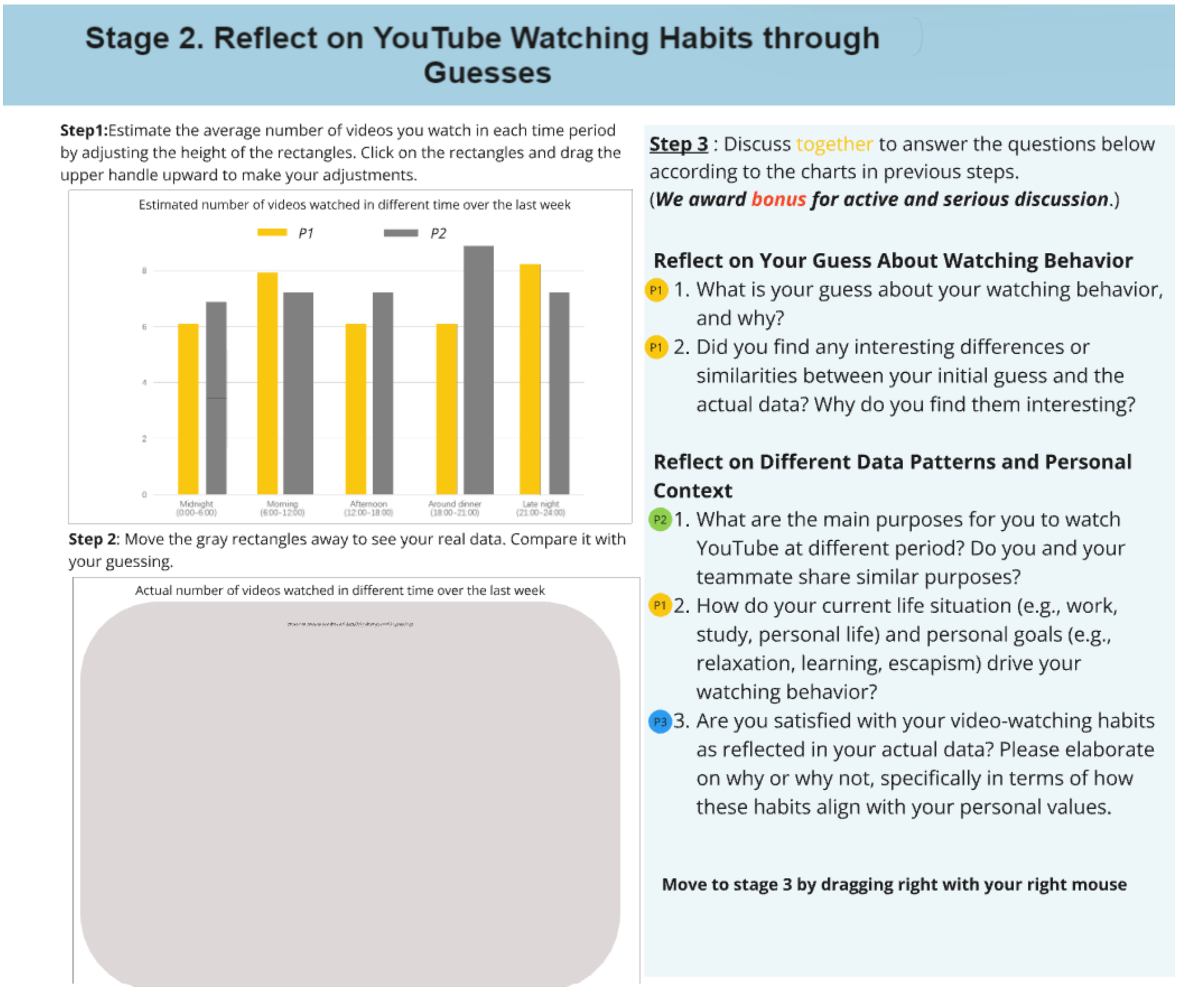


Figure A.2: Stage 2

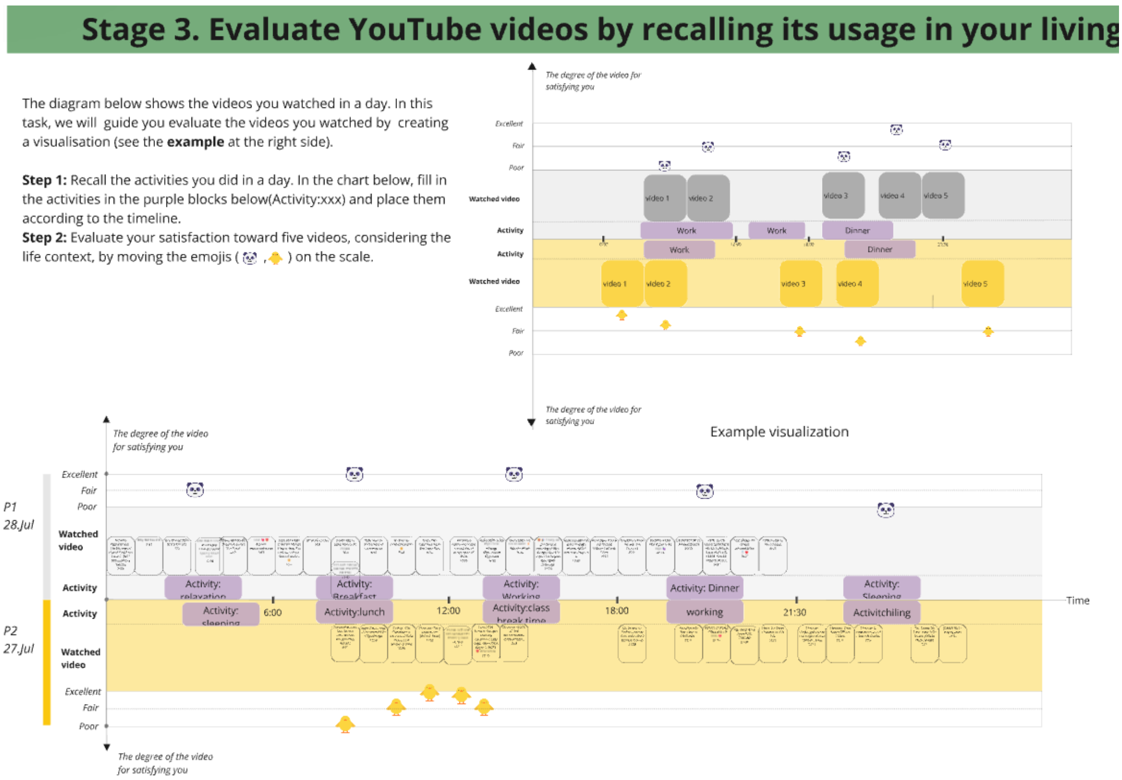
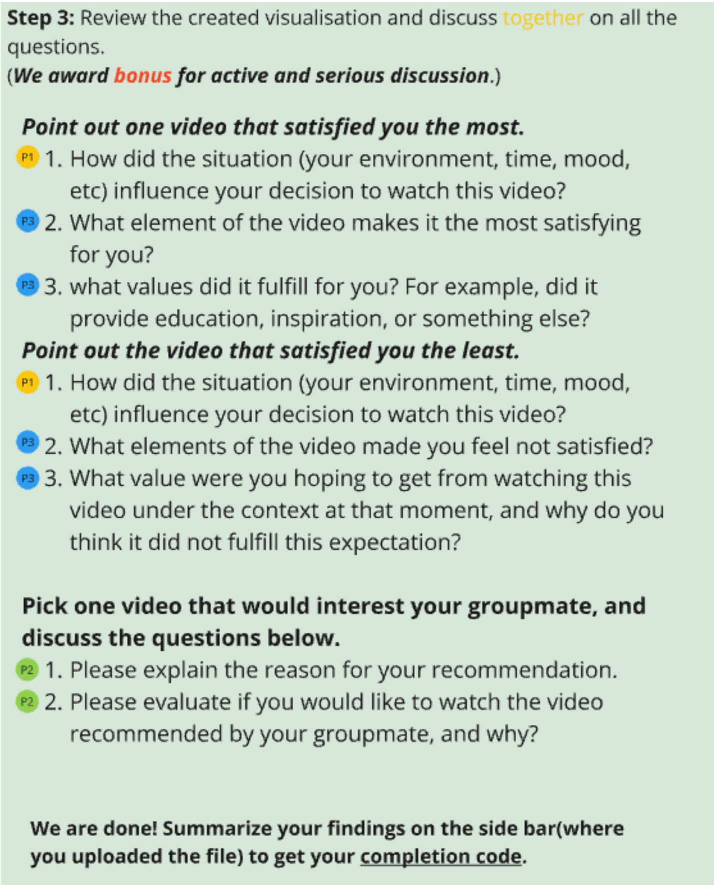


Figure A.3: Stage 3-1



Step 3: Review the created visualisation and discuss together on all the questions.
(***We award bonus for active and serious discussion***.)

Point out one video that satisfied you the most.
P1 1. How did the situation (your environment, time, mood, etc) influence your decision to watch this video?
P3 2. What element of the video makes it the most satisfying for you?
P3 3. what values did it fulfill for you? For example, did it provide education, inspiration, or something else?
Point out the video that satisfied you the least.
P1 1. How did the situation (your environment, time, mood, etc) influence your decision to watch this video?
P3 2. What elements of the video made you feel not satisfied?
P3 3. What value were you hoping to get from watching this video under the context at that moment, and why do you think it did not fulfill this expectation?

Pick one video that would interest your groupmate, and discuss the questions below.
P2 1. Please explain the reason for your recommendation.
P2 2. Please evaluate if you would like to watch the video recommended by your groupmate, and why?

We are done! Summarize your findings on the side bar(where you uploaded the file) to get your completion code.

Figure A.4: Stage 3-2