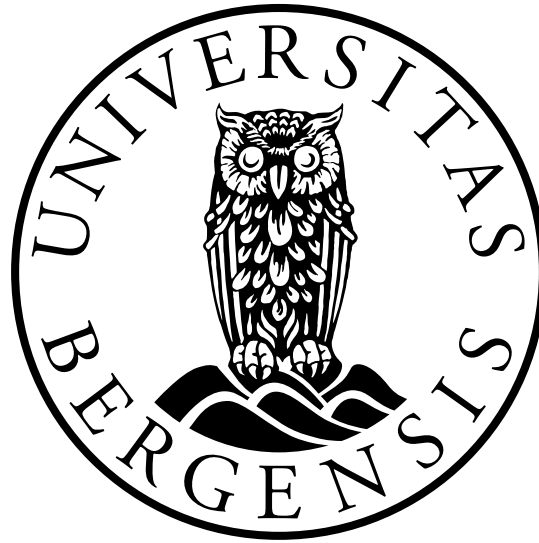


UNIVERSITY OF BERGEN



Department of Information Science and Media Studies

MASTERS THESIS

---

**Divide & Control:**  
**Evaluating Controllability in Multi-lists**  
**Presentation of Recommendations**

---

*Author: Johnny Bjånesøy*

*Supervisor: Assoc. Prof. Dr Christoph Trattner*

December 1, 2020



# Abstract

In recommender systems, the concept of control is associated with ways users can manipulate the system through interactions or by defining parameters in order to be provided more personal and better recommendations. Other studies in the movie domain have found that users may have a divergent perception of similarity regarding which features are important to them when looking for similar movies. This thesis sets out to investigate if these divergent opinions on similarity can be leveraged by controllability in the multi-lists presentation of recommendations. This thesis shows that user control did not appear to be evaluated more positively than a non-control recommender system for the average participant. This thesis found that multi lists presentation of recommendations without control were generally evaluated better than with control by performing a quantitative conditional user evaluation of the recommender system. When looking at participants' demographic properties, it may be that some subgroups consisting of users with a higher level of domain knowledge or similar system experience may favor control. Furthermore, no significant variances between the three list sort methods that the system uses to enforce the users' control were discovered. As controllability in recommender systems have not been extensively evaluated in the research corpus, this thesis hopes to be a starting point that can inspire future studies to attempt other novel approaches in implementing and evaluating controllability in the multi-lists presentation of recommendations to achieve more positive results.





# Acknowledgment

I would first thank my supervisor Assoc. Prof. Dr. Christoph Trattner for his guidance, support, and inputs throughout this entire thesis project. My thanks also goes to Alain Starke for his eager help and feedback whenever I was stuck with a problem. A thanks is also given to my fellow students for their willing participation during development and late night discussions on topic, and a special thanks goes out to Bjørn Helge Sandblåst for editorial support.

I am very grateful for the grant I received from the DARS lab (<https://dars.uib.no/>), which provided the financial means to run this study and reach the critical mass of participants. A hearty thanks to TV2 for awarding me a student scholarship that funded the the development and infrastructure of the study.

Bergen, Norway, November 2020

Johnny Bjånesøy



# Contents

<b>Abstract</b>	<b>ii</b>
<b>Acknowledgment</b>	<b>v</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation . . . . .	1
1.2 Problem . . . . .	2
1.3 Research Questions . . . . .	3
1.4 Contributions . . . . .	4
1.5 Thesis Outline . . . . .	5
<b>2 Background</b>	<b>7</b>
2.1 Multi-list Presentation of Recommendations . . . . .	7
2.2 Control Elements . . . . .	11
2.3 Demographics and Recommender Systems . . . . .	16
2.4 Summary and Key Differences . . . . .	19
<b>3 Methodology</b>	<b>21</b>
3.1 Prototype . . . . .	21
3.2 Research Design . . . . .	27
<b>4 Results</b>	<b>43</b>

---

4.1	Evaluation of Control Elements (RQ1) . . . . .	44
4.2	Evaluation Variances by Interaction Preference (RQ2) . . . . .	47
4.3	List Sort Methods Impact on System Satisfaction (RQ3) . . . . .	49
4.4	Demographic Effects on Evaluation (RQ4) . . . . .	51
<b>5</b>	<b>Discussion, Summary &amp; Future Work</b>	<b>57</b>
5.1	Discussion . . . . .	57
5.2	Limitations . . . . .	60
5.3	Future Research . . . . .	61
5.4	Open Science . . . . .	61
	<b>Bibliography</b>	<b>62</b>
<b>A</b>	<b>Screenshots</b>	<b>73</b>
<b>B</b>	<b>Statistics</b>	<b>79</b>
B.1	RQ1 . . . . .	79
B.2	RQ2 . . . . .	89
B.3	RQ3 . . . . .	92
B.4	RQ4 . . . . .	94

# List of Figures

1.1	A general overview of the thesis structure. . . . .	5
2.1	Nanou et al. structured overview of recommendations system . . . . .	8
2.2	Viaplay multi-lists presentation of recommendations . . . . .	10
2.3	The Tasteweight recommender system . . . . .	14
2.4	The Setfusion recommender system . . . . .	15
2.5	Feature importance rating by users . . . . .	16
2.6	Demographic distribution in commonly used open source datasets . . . . .	17
3.1	The developed controllable multi-lists recommender system . . . . .	26
3.2	Flowchart and screenshots providing an overview of the study . . . . .	28
3.3	Overview of the sample distribution per condition of this study. . . . .	31
3.4	Process chart highlighting the operations of no control (A). . . . .	34
3.5	Process chart highlighting the operations of control (B). . . . .	35
4.1	List order satisfaction plot chart . . . . .	50
A.1	Instruction screen screenshot . . . . .	74
A.2	Search screen screenshot . . . . .	75
A.3	Instructions for control screenshot . . . . .	76
A.4	Recommendation interface screenshot . . . . .	77
A.5	Survey screenshot . . . . .	78

---

B.1	System usability metrics errorbar plot RQ1 . . . . .	80
B.2	Activity metrics errorbar plot RQ1 . . . . .	81
B.3	Recommendation quality metrics errorbar plot RQ1 . . . . .	81
B.4	System satisfaction metric errorbar plot RQ1 . . . . .	82
B.5	Demographic metrics errorbar plot RQ1 . . . . .	82
B.6	System usability metrics errorbar plot RQ2 . . . . .	89
B.7	Activity metrics errorbar plot RQ2 . . . . .	90
B.8	Recommendation quality metrics errorbar plot RQ2 . . . . .	90
B.9	System satisfaction metrics errorbar plot RQ2 . . . . .	91
B.10	Demographic metrics errorbar plot RQ2 . . . . .	91
B.11	Errorbar plot RQ3 . . . . .	93



# List of Tables

2.1	Commercial evaluation results	12
3.1	Dataset attributes	22
3.2	Used similarity functions	23
3.3	Similarity functions overview	24
3.4	Overview of conditions and names.	29
3.5	Questionnaire overview	37
3.6	Overview of the tracked activity metrics used in this study.	38
3.7	Sample size overview for the conditions.	40
4.1	Result of t-test between A-B	45
4.2	Result of t-test between interaction elements	48
4.3	List Satisfaction Comparison Table	49
4.4	Sample overview	51
4.5	Results from system experience influence analysis RQ4	52
4.6	Post hoc results on age	54
4.7	Post hoc results on domain experience	55
B.1	Result of t-test between A1-B1	84
B.2	Result of t-test between A2-B2	85
B.3	Result of t-test between A3-B3	86



---

B.4	Post hoc results from Table B.5 . . . . .	87
B.5	Interaction effect analysis result . . . . .	88
B.6	Results from gender influence analysis . . . . .	95
B.7	ANOVA result on age and domain experience . . . . .	96



# Chapter 1

## Introduction

### 1.1 Motivation

The popularity of digital streaming platforms for movies has significantly increased through the last decade, catalyzed both by the increased availability and ease-of-use for consumers when attempting to find content[22]. The last decade has seen the rise of digital streaming platforms for movies, made popular by the ease of finding content to consume for users. However, as these platforms have grown in niches and content, this growth starts to complicate the user experience. A Netflix study reveals that users on their site lose interest after either 60-90 seconds of idle browsing or investigating 10 to 20 movies [22, 3]. Maintaining user interest while still providing a substantial library and recommendation section is a central point of interest for all parties.

Historically, this consists of improving the accuracy and precision of systems [46]. This approach can be problematic in a subjective domain as movies where users may have different perceptions of similarity and disagree on topics such as what a genre would constitute. A study on comparing humans versus algorithms found that recommender systems performed best in general [34]. However, when the tested user had niche preferences, humans severely outperformed the algorithms [34]. Several other studies have revealed that users' have different ideas on what features are best for similarity calculations [54, 55, 57, 53] and conflicting results on what methods are best at recommending [13, 12]. The justification for this topic lies in these reported discrepancies between how recommender systems perceive similarities and how users do it.

An interesting area of exploration is investigating if these noted effects could be leveraged to make recommender systems better. A potential scenario this could be leveraged in is a controllable multi-lists presentation of recommendations, which is the organization of recommendations into a structured list based on features or other grouping methods as used in commercial streaming services such as Netflix [5, 22, 4]. Such multi-lists presentation of recommendations is not uncommon for commercial services. However, it is restricted to providing a general overview of content or homepage and is rarely used in similar item recommendations, as revealed in a commercial evaluation study in this thesis. Multi-lists presentation of recommendations is suggested by literature to be of use in the context of a similar item recommendation scenario [11] and suggested to possible to be improved with control [31].

## 1.2 Problem

The focal point of this thesis is whether user control of recommendation features can be beneficial or not in the movie domain. The problem is if giving users control over a multi-lists presentation of recommendations in a similar item recommender system would improve their experience with the system. While multi-lists is often used for a general overview of content in a streaming service by grouping recommendations into categories [22, 4, 31, 5], the context is set to a similar item recommendation scenario where users are looking for recommendations to a given movie. The selected approach is underexplored both commercially and in research literature [31, 11]. No similar research has attempted to evaluate if control in a multi-list presentation of recommendation is beneficial to the author's knowledge. Consequently, the overall problem this thesis explores is summarized as follows:

*Does controllability in the multi-lists presentation of recommendation in a similar item recommendation scenario improve the user experience?*

## 1.3 Research Questions

Based on the general problem statement, four research questions focus on different aspects of this problem.

- **RQ1:** *To what extent does controllability in a multi-lists presentation of recommendation for similar item recommenders improve the users' experience?* Section 4.1 details pairwise comparison analyses between multi-lists recommendation interfaces with control and without control performed to see variances in users evaluation.
- **RQ2:** *How do different interaction methods for control affect participant evaluation?* Based on the literature, two interaction methods were selected for control: drag & drop and clickable arrow buttons. By grouping participants based on which interaction method they primarily used for control, Section 4.2 details comparison analysis between these groups performed to see if interaction methods impacted the evaluation.
- **RQ3:** *To what extent do different list sort methods for user control impact the user evaluation of the recommender system?* In Section 4.3, different list sort methods based on the users' control of the multi-lists recommendations are compared to see variances in evaluation based on how the system interpreted user control in terms of list sort methods.
- **RQ4:** *To what extent do demographic properties and familiarity with similar systems affect system evaluation?* In Section 4.4, the data are divided into groups for statistical analysis to see if any demographic properties are influential to participants evaluation of the recommender system.

## 1.4 Contributions

The thesis provides the following contributions to the research corpus:

- Insight into how users would interact with such a system (RQ2). Two interaction methods were enabled for control in terms of clickable arrow buttons and drag & drop. Results show that one-third of the participants never interacted with the system, with only 13% of all users showing a preference for drag & drop.
- The thesis evaluated three different list sort methods for controllable multi-lists presentation for recommendations (RQ3). The analysis did not find any significant variance between the methods, signifying that the list sort method may not play a major role in user evaluation of such a system.
- Several demographic metrics in similar research were found to impact user evaluation of recommender systems (RQ4). Evaluating these in the context of this thesis found age and gender to have little to no influence. However, it was found that user experience with similar systems and domains appears to impact their evaluation of this system.
- Lastly, the data from the study performed in this thesis, consisting of 300 participants and 140 metrics, are made available for future research. This data consists of both the raw JSON files with complete data and cleaned CSV files with the subset of this thesis's metrics. A pipeline was created to create CSV files, which can easily be extended to include additional metrics. In addition, the prototype code is made available. All this is made open source, detailed in Section [5.4](#).

## 1.5 Thesis Outline

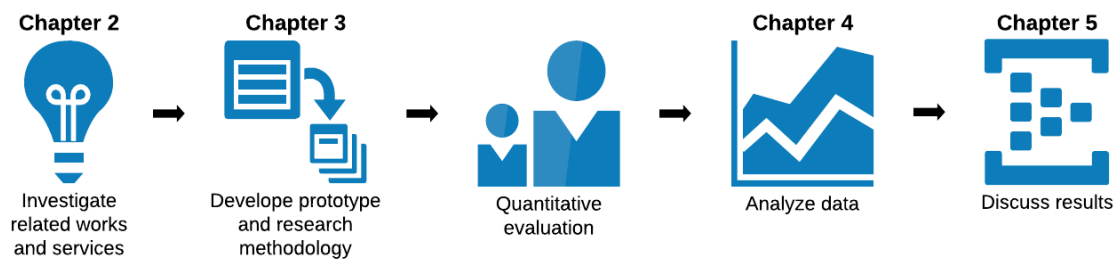


Figure 1.1: A general overview of the thesis structure.

The general outline of this thesis, as visualized in Figure 1.1, is as follows:

- **Background.** Chapter 2 details the literature found on the three focal points of this thesis: multi-list presentation of recommendations in Section 2.1, evaluation of commercial services and research on control in Section 2.2, and demographic evaluation in recommender systems in Section 2.3.
- **Methodology.** Chapter 3 describes the prototype development in Section 3.1 which entails the data, similarity functions, and details on the implemented control. Section 3.2 elaborates on the methodology concerning the study design, metrics, and how this relates to providing results for the research questions.
- **Results.** Chapter 4 presents the results from the statistical analysis performed to address the research questions.
- **Summary & Conclusions.** Chapter 5 discusses and summarizes the findings concerning the research questions postulated based on the results from the previous chapter. This chapter also entails the limitations of this study, future research suggestions, and open science.





# Chapter 2

## Background

The background chapter provides an overview of previous work relevant to this thesis and is divided into four sections.

- Section 2.1 investigates studies on multi-lists presentation of recommendations.
- Section 2.2 details relevant research on control in recommender system. Additionally, an evaluation of control and presentation element in commercial movie streaming services are detailed.
- Section 2.3 investigates which demographic properties that may influence the study in this thesis based on literature.
- Section 2.4 concludes the chapter and elaborates on the critical differences between the research discussed and this thesis.

### 2.1 Multi-list Presentation of Recommendations

One of the primary purposes of recommender systems is to reduce the choice overload that comes with selecting from an extensive catalog [44, 7]. The resulting overload often leads to the paradox of choice, in which user satisfaction is often higher when the user has a choice between a few items rather than many [43]. Scheibehenne et al. [49] performed a meta-analysis on several previously performed experiments on choice overload and found non-reproducible results and an average effect across all evaluated experiments of zero,

with many conflicting results concerning user satisfaction when interacting with long lists. Scheibehenne et al. [49] note in their study that the choice overload appeared to be present, but no causal link could be established as to what exactly causes this effect [49]. Bollen et al. [7] followed this study to analyze the effect in the movie domain, testing out different lengths of top-n recommendations with no evaluation difference found between the list length. Bollen et al. argue that while a more extensive list may demand a higher cognitive load for a decision, it also increases satisfaction when used as the recommendations are more diverse, proposing that a constant tradeoff between cognitive load and utility exists that stays somewhat neutral as the list size changes [7]. What may be surmised from this is that choice overload is not just about how much information is presented to the user, but the presentation, the users' own goals, and the tools available to them may tip the balance between utility and cognitive load to either be beneficial or detrimental to user satisfaction.

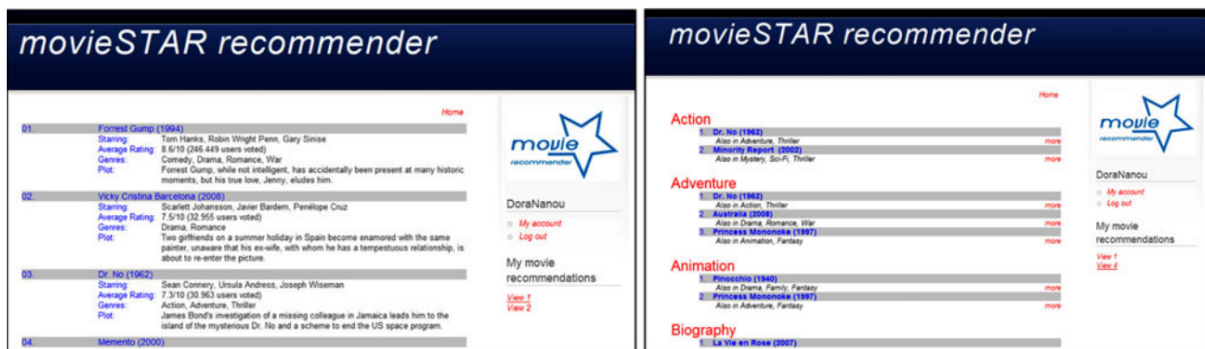


Figure 2.1: Excerpt from Nanou et al. study [41], top-N(left) presentation of recommendations versus structured overview of recommendation results(right), which was evaluated to be better by participants.

A typical recommender system has the end goal of producing a top-N list of recommendations for a user [46, 32, 30], regardless of how many different features and algorithms it takes into consideration [46, 32, 30]. There have been suggestions of splitting such top-n lists into multiple lists, often denoted as a *multi-lists* presentations based on increased diversity [31], user satisfaction [52] and ease of browsing [5]. Some insight into why this may be preferable is presented by Alvino & Basilico of the Netflix research team that elaborated in a blog post on why they use a multi-lists presentation of recommendation for their home page [3]. Their evaluation is that users only need to look at a few entries and the row label to decide whether they wish to investigate the current row further or move on. The organized row display may be considered a more intuitive way to quickly browse larger item segments than a simple

top-n list or grid [22].

Several studies found have evaluated and experimented with multi-lists presentation of recommendations. One such study by Nanou et al. [41] looked at recommendation presentation utilizing the movieSTAR framework in which different visualization strategies were employed. One of the favorable presentation from their study was what they coined *structured view*, in which recommendations were organized by genre and evaluated positively by users [41], visualized in Figure 2.1. Pu et al. [29] performed a study on the amazon store where top-N recommendations were compared to dividing the recommendations into different tabs, based on similar users, price, content, or popularity. The results were that users had a higher preference for the tab-split recommendation presentation than the top-N presentation. Participants also noted a higher level of diversity in the tab-split interface, even when the content of both interfaces contained the same level of measurable diversity [29]. This can lead one to conclude that the presentation may affect how users perceive diversity or evaluate more items. Pu & Chen also published a paper reviewing the literature to postulate possible guidelines for the development of recommender systems, which includes multi-list presentation [11]. Another study is the *RealCode* framework, which, among other historical and social features, splits the top-N recommendation into chronological lists, allowing users to browse through release years and look at previously historically viewed movies [10].

Having a multi-lists presentation of recommendation does add a new layer of complexity to the recommendation process by needing a function to rank and order the lists. Concerning RQ3, it of interest to investigate different ways to interpret the users' control over the multi-lists presentation in terms of list sort methods. However, no similar research investigated this problem in conjunction with the multi-lists presentation, so the focus on similar research was to look at the literature on different ranking similarity functions and methods in a hybrid top-N recommendation list. Such topics may include weight application [55], user algorithm selection [12, 9], or content sorting [36, 21], which, while not a multi-list presentation in itself, does lend some knowledge transfer in regards to shared goals and methods. An example of this is a content-based system that utilizes feature weight implicitly learned from the features of a movie that a user has consumed to predict new recommendations [55]. Another identified approach is to use defeasible argumentation to adjust weights and algorithms for a given user, based on previous history [9].

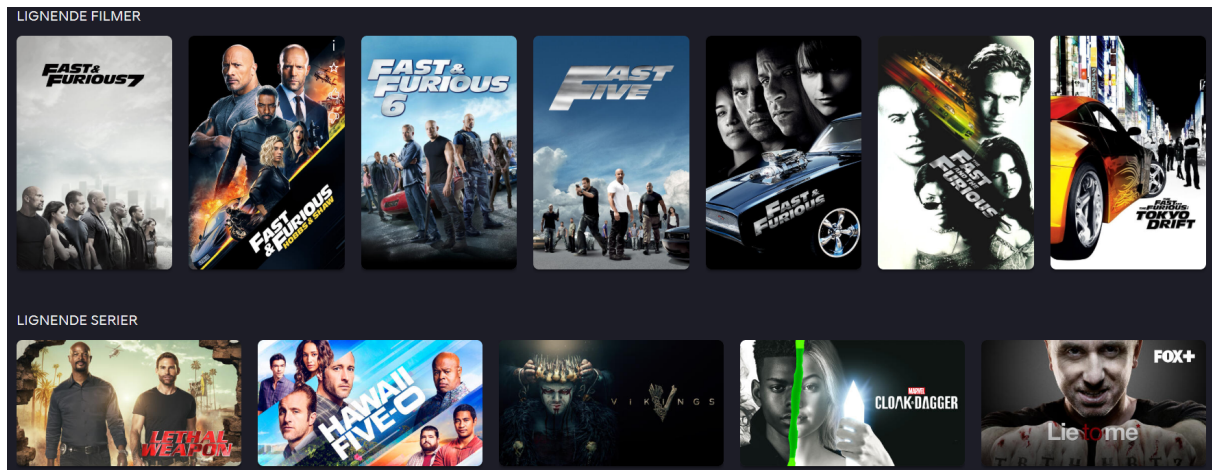


Figure 2.2: How Viaplay presents recommendations as multi-lists in a similar recommendation scenario. Screenshot from 24th of August 2020.

While some commercial streaming services are already utilizing similar presentation [22] for their general browsing page, an evaluation carried out by this study as seen in Table 2.1 found only one service that provides a multi-list presentation of recommendation in a related item context. The service was Viaplay<sup>1</sup> which split the recommendations in a similar item scenario into movies and series when possible.

The most similar study found in terms of controllable multi-lists presentation of recommendation is an experiment ran by Ekstrand et al. on the MovieLens platform. In this experiment, users could choose which algorithm to provide them with recommendations [20]. This experiment was based on a previous user evaluation study on different methods [19]. Participants were randomly assigned to one of four recommendation methods with non-descriptive names, which hid their functionality. Two of the algorithms were non-personalized as most popular and random, while two were based on collaborative filtering. Participants assigned to the non-personalized ones switched the most, and participants generally switched between a few before settling, fairly divided between the two CF methods [21]. It is noted that only 25% of the participants ever made a switch [20]. Based on these studies, a follow-up paper argued for user control in algorithm selection and how this could improve both satisfaction and accuracy in systems based on results from this and other studies [18].

To summarize, several studies find the multi-lists presentation of recommendation an improvement over one list top-N recommendation presentation [45, 41, 51, 31, 10, 11] and sev-

<sup>1</sup>Viaplay is a Nordic streaming service by Viasat. [www.viaplay.no](http://www.viaplay.no)

eral studies that find positive effects in feature weight adjustments [55, 9], no commercial services in the movie domain utilize them. The most interesting missing feature in all of the studies investigated is that none of them includes control in their multi-lists. While some postulate that control has utility [31, 11], no studies investigated has set out to evaluate this through prototyping and user evaluation. Since no similar studies are found that have evaluated control in multi-lists, we are forced to look into the general domain of recommender systems to gather more knowledge on implementation and evaluation of control.

## 2.2 Control Elements

Control in the context of recommender systems are elements that enable users to manipulate how their recommendations are presented or by providing feedback that the system comprehends [31]. Most recommender systems rely on behavioral data such as click-through rate or by users rating items [46, 31, 30], which offer users some small control in terms of direct feedback. However, more elements and aspects of recommendation are possible to be controlled. Reorganization, setting parameters, and exclusions are examples of direct control elements in a recommender system. As preparatory work for the thesis, an evaluation study in commercial streaming services was conducted. The goal was to uncover what kind of control and presentation elements are currently used in commercial services. Services were selected based on their availability in Norway as either international or national. This survey was carried out on the 24th of August 2020.

As seen in Table 2.1, control is rarely a feature in commercial services. None of the elements utilized in this thesis evaluation study detailed in Chapter 3 are present. Why control is rarely enabled in these services is unknown and can only be speculated on as the literature argues that control elements are potential of high utility [18, 25, 8, 45, 2]. While some of the services displayed ratings from either IMDb or Rotten Tomato<sup>2</sup>, none provided users the ability to rate movies in their services outside of the binary system in Netflix and Hulu, which may be explained in a Cosley et al. study [14].

---

<sup>2</sup>IMDB and Rotten Tomato are communities for rating and critiquing movies.

Table 2.1: Evaluation results from commercial streaming services regarding which control and presentation elements are present in their systems. Performed 24. August 2020.

Service	Likert	Binary	Remove	Rearrange	Seen	M-List
Netflix <sup>3</sup>		✓				
TV2 Sumo (NO) <sup>4</sup>					✓	
Dplay (NO) <sup>5</sup>						
HBO Nordic <sup>6</sup>						
NRK TV(NO) <sup>7</sup>						
Viaplay(NO) <sup>8</sup>						✓
Prime Video <sup>9</sup>						
Strim(NO) <sup>10</sup>						
Blockbuster <sup>11</sup>						
Disney+ <sup>12</sup>						
Hulu <sup>13</sup>		✓	✓			

**Likert:** Users can rate movies on a Likert-scale.

**Binary:** Users can rate movies in binary terms.

**Remove:** Users can remove a recommendation provided.

**Rearrange:** Users can rearrange the recommendations.

**Seen:** Users can mark a recommendation as already seen.

**M-List:** Service has a multi-lists presentation of recommendations in a similar item scenario.

To begin discussing related research on control in recommender systems is a McNee et al. study on the cold start problem [39]. A typical solution that is sometimes used on this problem is to task a new user with rating a certain amount of movies before letting the user use the service [30, 46]. In this scenario, McNee et al. compared users who had to manually search for movies to rate with users who were provided a random set of movies suggested by

<sup>3</sup>Netflix: <https://www.netflix.com/>

<sup>4</sup>TV2 Sumo: <https://sumo.tv2.no/>

<sup>5</sup>Dplay: <https://www.dplay.no/>

<sup>6</sup>HBO Nordic: <https://no.hbonordic.com/>

<sup>7</sup>NRK: <https://tv.nrk.no/>

<sup>8</sup>Viaplay: <https://viaplay.no/>

<sup>9</sup>amazon Prime Video: <https://www.primevideo.com/>

<sup>10</sup>Strim: <https://www.strim.no/>

<sup>11</sup>Blockbuster: <https://blockbuster.no/>

<sup>12</sup>Disney+: <https://www.disneyplus.com/>

<sup>13</sup>Hulu: <https://www.hulu.com/>

the system to rate [39]. The result was that users had a higher preference for searching for movies to rate in opposition to being given a selection, even if the recommendation's initial accuracy was lower in comparison, and the temporal and cognitive cost increased [39]. This points towards the possibility that users value more control rather than the system having more accuracy.

One control element is recency and popularity modifiers. Harper et al. [25] conducted a study utilizing arrow buttons that adjusted movie recommendations in a top-N list. No information was given to the participants on the buttons' functionality, but secretly adjusted the recency or popularity weight on the recommendations [25]. Participants were tasked with adjusting the list with these buttons until they were satisfied with the recommendations. The results were that participants not only utilized these buttons but had wildly different parameters set when they evaluated their list to be at its optimal stage [25]. The results show that users may have different opinions on what features are best; some may enjoy popularity while others may be unfamiliar with newer movies and prefer old ones. Commercially, this type of control element is also utilized in the video game sales platform named Steam, albeit as sliders instead of buttons<sup>14</sup>.

---

<sup>14</sup>Utilized in the discovery section of the Steam video game platform. [www.steampowered.com](http://www.steampowered.com)



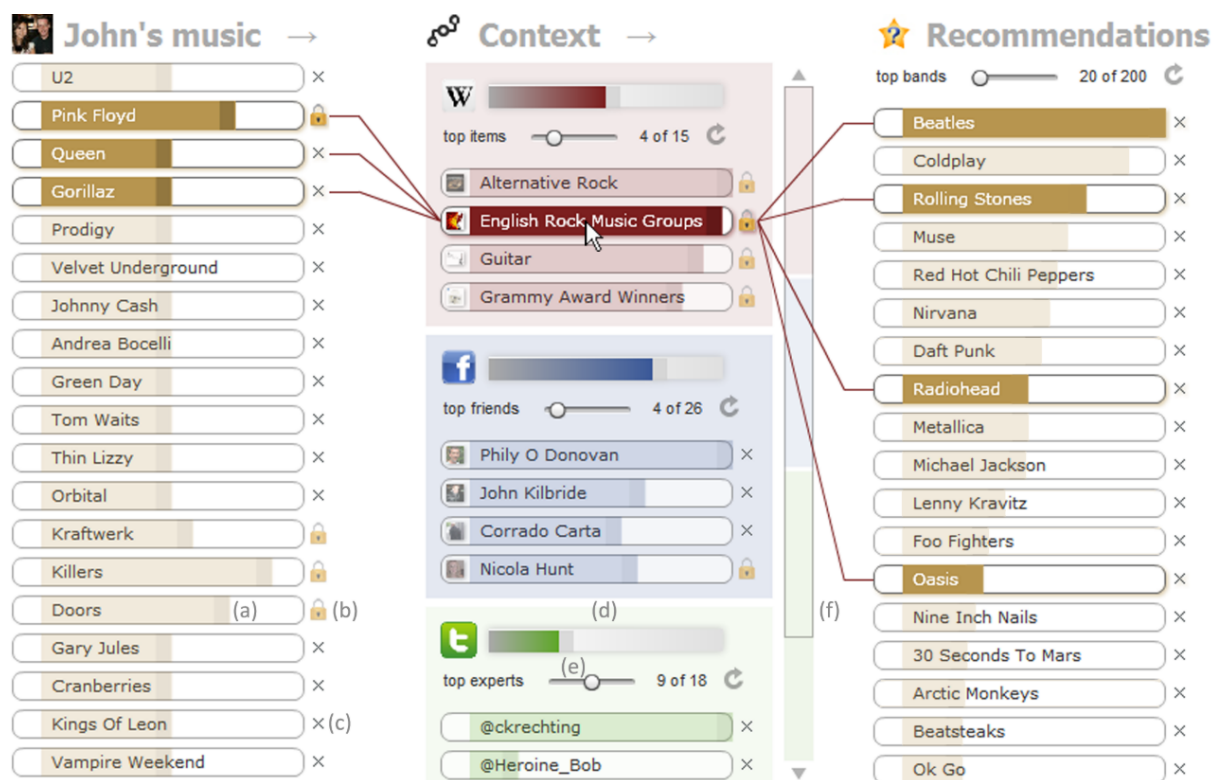


Figure 2.3: The Tasteweight recommender system control interface from the Bostandjiev et al. study [8].

One extensive framework on music recommendation called *Tasteweight* [8] allows users to control several different parameters of the recommendation process as seen in Figure 2.3. The evaluation study on this framework reported that users enjoyed control with a positive impact on accuracy and user satisfaction [8]. Another framework found is the *uRank* framework for literature search, which allowed users to set parameters from keywords through drag & drop [16]. The available pool of keywords was suggested by the system based on occurrences in the reference document or custom queried from the user [16]. An alternative framework in the same domain is *SetFusion* by Parra [44]. This tool for literature recommendation provides users with the ability to adjust feature weights using sliders. It presents the results in a Venn diagram, which allows users to assess the relevancy of the articles [44], displayed in Figure 2.4. Keyword weight and frequency could also be adjusted through sliders. Kveton & Berkovsky [36] utilized similar filtering control elements by users clicking on tags to prune their recommendation list in their study on minimal interaction recommendation system [36].



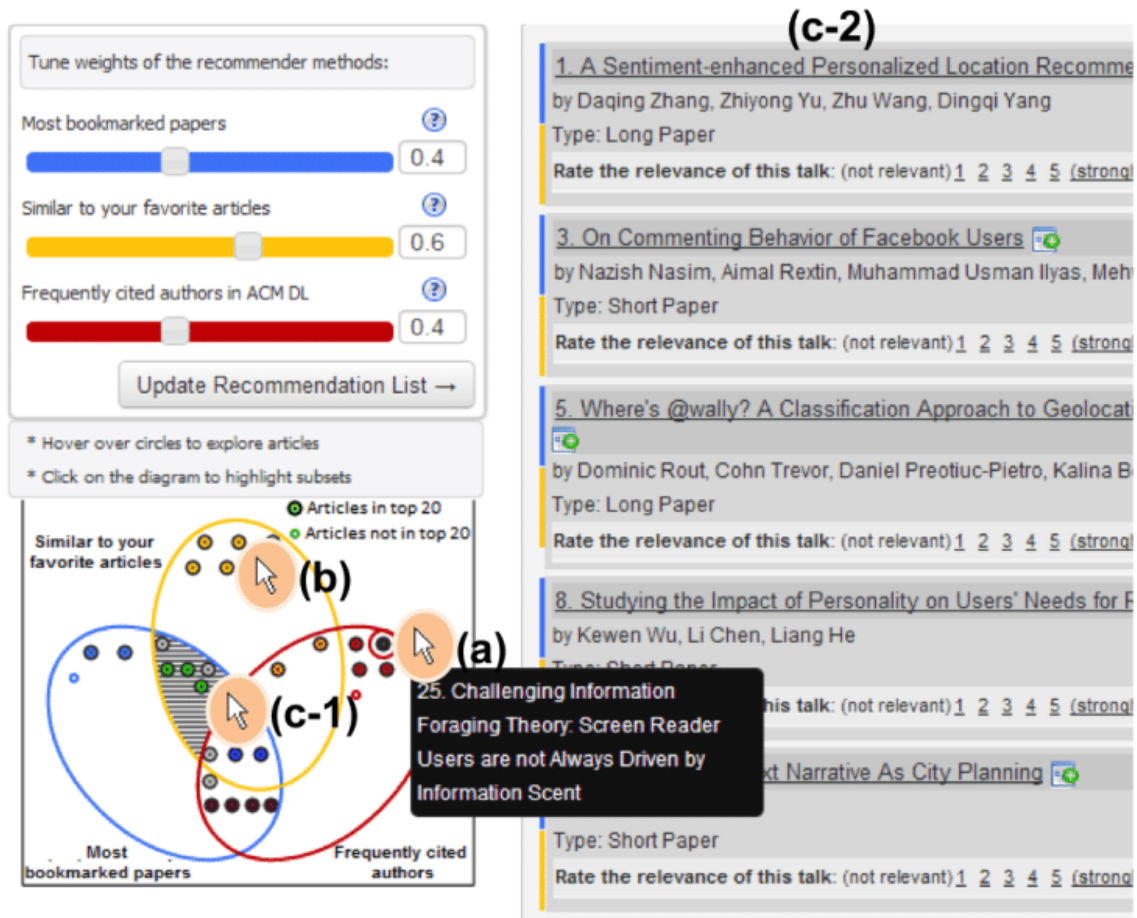


Figure 2.4: Sefusion framework for literature recommendation visualization taken from Parra study [44]. Users could control their recommendations by adjusting weights and saw recommendations organized in a Venn diagram.

A counterpoint to enabling controllability is that users may not have a correct understanding of what factors influence them. Odic et al. [42] tasked users with evaluating how they thought different factors influenced what movie they were going to watch next and compared this with statistical testing on rating data. The results found only a low level of overlap. An example is that users evaluated that weekdays had little impact on their movie selection, while the statistical analysis revealed it to be a significant factor [42]. Thus, too much control may be detrimental, as users do not necessarily have an accurate understanding of their decisions. Divergent opinions are also noted by Yao & Harper [57] who tasked participants of their study to rate what features mostly impacted their decision on what to watch next. The results displayed in Figure 2.5 shows that while some factors are generally important for all, divergent opinions may imply that different users have different needs and goals when interacting with a recommender system [57].

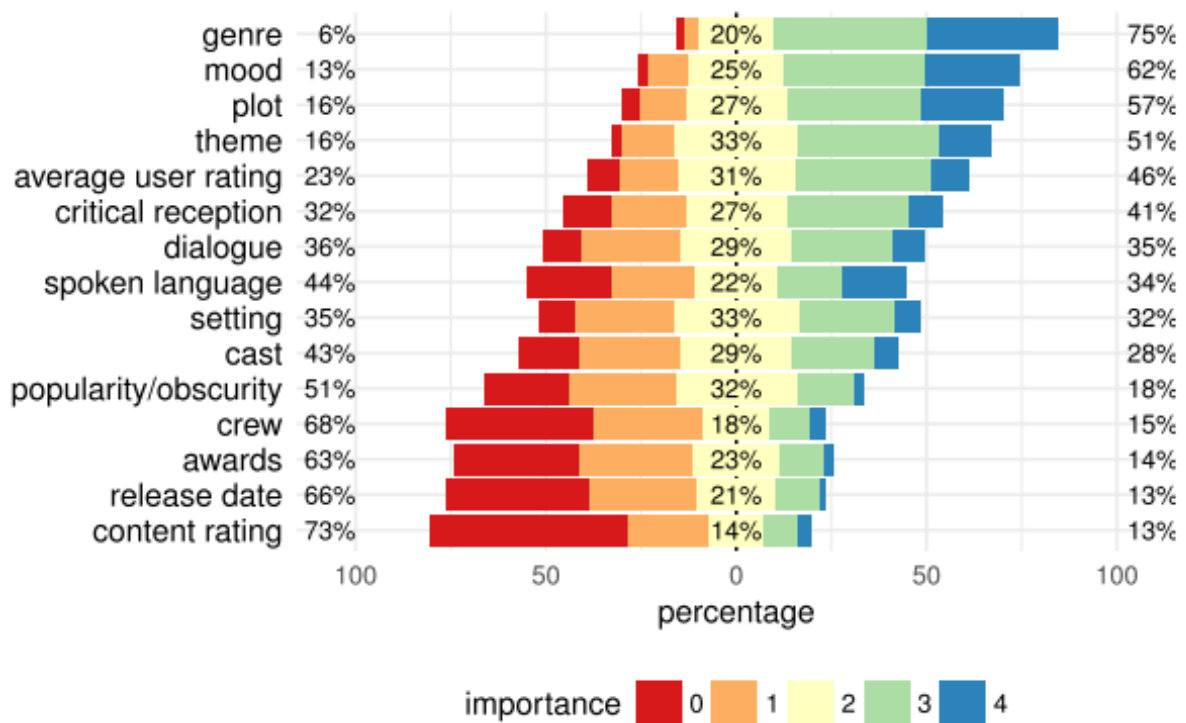


Figure 2.5: Feature importance results taken from Yao & Harper study [57]. Rating given by participants in the study on feature importance when deciding on what movie to watch.

The preceding research highlights how several different interaction methods such as sliders, buttons, and drag & drop have been used for control in recommender systems, which is of interest to RQ2. While no experiments or frameworks on controllable multi-lists presentation as investigated in this thesis was found, control elements appear to be beneficial to a recommender system [18, 25, 8, 45, 2]. However, specific examples need to be evaluated to verify their utility.

## 2.3 Demographics and Recommender Systems

Another area of interest is investigating if users' demographic properties may affect their evaluation or use of the system, as per RQ4. This section goes into similar research and experiments uncovered that relate to demographic topics in recommender systems.

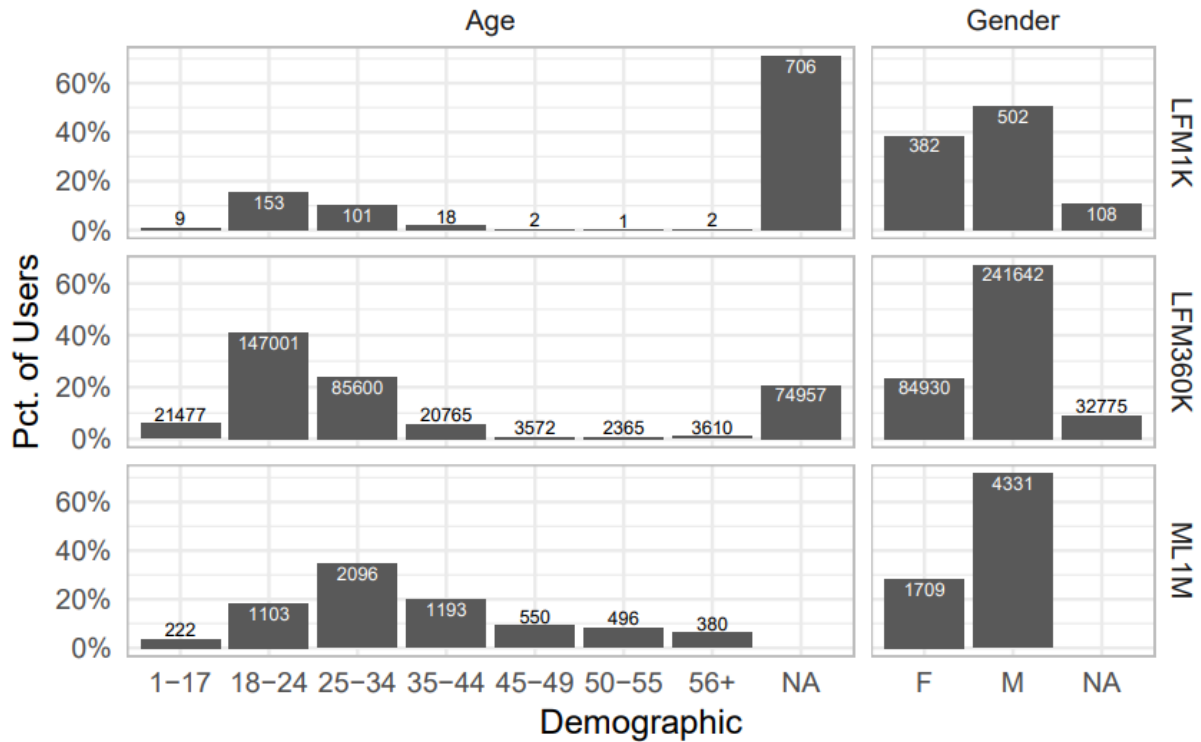


Figure 2.6: User demography of datasets taken from Ekstrand study [21] that highlight demographic distribution in commonly used datasets in research. LFM1k and LFM360k is two published datasets from the last.fm music community, ML1M is a published movieLens dataset.

A study by Ekstrand et al. in 2018 looked into recommender accuracy and user retention rate in services such as the music platform of last.fm and the MovieLens movie recommendation platform. Openly available datasets from these platforms were often used by researchers and analyzed for their demographic distribution [21]. This was achieved by looking at the demographic data behind the ratings and consumption and running offline evaluations with nDCG. As displayed in Figure 2.6, there is a clear overweight in the movieLens dataset ML1M of both young adults and males. The median consumption is also slightly higher per user of these segments [21] and was evaluated to have a higher retention rate and nDCG. Reducing the popularity bias of the data also had an impact on reducing demographic evaluation differences, though while the effect is still somewhat noted in all data sources, a causal link was not established [21].

Outside of evaluating the data, the MovieLens demographic data can be used for setting up demographic recommender systems. In a study by Al-Shamri [1], different user profiling methods that utilized demographic data were evaluated. It is noted in the findings that while there are benefits of utilizing each of the demographic properties alone, they work best in

unison to compensate each other [1]. Other studies on the subject also found a demographic discrepancy in the results. A study on a German dataset of movies found that gender had a high positive impact on precision when highlighting similarities to other same-sex users, but little in terms of city and age [48]. In contrast, a study investigating Docear, a German research literature recommendation service, found that the click-through rate on recommendations increased significantly with age [6]. The click-through rate also increased with the number of recommendations provided to the user and how often a user had previously used the system.

While there is a lack of studies that evaluate the results directly based on demographic data, some studies infer or are directly provided demographic data to increase recommendation accuracy [46, 30]. The value is that if including age as a factor in recommendation increases accuracy, then age may be implied to be a factor. One study that investigated this is a study by Sun et al. [50] that investigated how to use demographic information to sort cold start problems while preserving privacy. They found that users who provided demographic information when starting to use their system had improved accuracy over users who withheld information [50].

In the domain of online cooking communities, a Rokicki et al. study was carried out to investigate if prejudices and assumptions regarding male and female cooking habits were factual [47]. Through data mining of online recipe sites, it was discovered that gender significantly impacted what recipes were investigated, the ingredients utilized, and the response given to authors. Additionally, using this in a simple recommender prioritizing same-gender data for a recommendation between a user and recipe author evaluated positively [47]. In terms of presentation and control, the Storytime framework for a book recommendation to children showed that specialized design and simple interaction elements were highly appealing to children [40] which may suggest that different recommender systems and frameworks appeal more to different demographic segments.

The research highlights how demographic properties are a factor that can impact recommendation when utilized either as a feature or when determining neighborhoods [1, 35, 56, 50]. Furthermore, demographic properties seem to affect both what type of content a user consumes and design preferences [40, 47, 21, 6, 48]. Based on the discussed literature, four properties seem to have some effect on participants evaluation and use of recommender

systems: Age [21, 6, 40], gender [48, 47], experience with similar systems [6, 35] and experience in the domain [6, 40]. However, results do differ between domains, so it is of interest to further investigate which are influential in the movie domain for RQ4.

## 2.4 Summary and Key Differences

The first prominent aspect is that controllability in the multi-lists presentation of recommendations is a completely novel approach with no similar research existing to the author's knowledge. Concerning RQ1, general implementations of control appear to be beneficial in other recommendation contexts [18, 20, 8, 44, 25, 31, 39, 36, 16, 11], with multi-lists presentation of recommendation being implied to be of high utility both from theoretical [45, 51, 52, 7, 31, 57] and from evaluation studies [41, 10, 29, 11]. For RQ2, no studies compared interaction methods for control, but some literature analysis set guidelines for implementation [11, 31, 51, 52]. For RQ3 and potential list-sort methods, weight ranking and persistent sort seem the most promising [57, 55, 12, 9, 54, 11]. Finally, for RQ4, the topic of how demographic properties may influence recommender system evaluation is not sufficiently explored, with some metrics selected for evaluation in this study based on similar research [21, 40, 47, 6, 48]. This thesis concludes from this background investigation that while no comparative studies have been performed on the subjects this thesis investigates, the literature supports the thesis from a theoretical standpoint and provides viability to the prototype developed for evaluation in Chapter 3.



# Chapter 3

## Methods

The methodology for evaluating the defined research questions is detailed in this chapter. This chapter begins with Section 3.1 that describes the controllable multi-lists presentation of recommendations in a similar item scenario prototype developed for this thesis. This section includes a description of the dataset, similarity functions, technical details, design, and the final prototype. With the controllable multi-lists recommender system described, Section 3.2 elaborates on the research design. It begins by providing a general overview of the quantitative conditional study before elaborating on the conditions selected, a step by step description of the study process for participants, metrics selected, and concludes with an overview of the statistical analysis methods.

### 3.1 Prototype

The prototype description begins by describing the data and similarity functions used to provide recommendations. The interface is then described in terms of the implementation of a controllable multi-lists presentation of recommendation. During the development, several quick and dirty qualitative trials [11, 28, 23] of the prototype were performed with participants from the local student population of the university. The primary goal was quality assurance and bug testing, but feedback on implementation was welcomed and impacted development.

### 3.1.1 Dataset

The utilized dataset stems from the research by Trattner & Jannach [54]. This is a modified version of the open MovieLens dataset<sup>1</sup> by GroupLens [24]. The version utilized is the latest full version updated 01.09.2018. The dataset is modified from the original full version by reducing the total number of movies from 58 000 to 2512 [54] by focusing on the most rated movies in the dataset. This pruning was performed to make sure the movies in the system are more familiar to the average participant to evaluate the prototype better. Such pruning actions are noted to be shared in similar research [57, 45, 53, 19]. Outside of the MovieLens dataset, the image covers of the movies are retrieved from The Movie Database<sup>2</sup>. The available attributes in the dataset used for similarity calculations can be seen in Table 3.1.

Table 3.1: Movie attributes available in the movieLens dataset for use in similarity measurements.

Attribute	Description
Title	Title of the movie.
Cover	Cover image of the movie (From TMDB).
Plot	Plot summary description.
Stars	List of main actors.
Genres	Classified genre(s).
Release Date	Date of first public release.
Ratings	Ratings data by movieLens members.
Director	Director(s) of movie.
Tag	Tags given by movieLens members.

### 3.1.2 Similarity Functions

In the preliminary research of Trattner & Jannach [54], similarity calculations were performed on the dataset described. These similarity calculations have with permission been appropriated for this study. An overview of these similarity calculations are viewed in Table 3.3.

Functions were selected to cover as many movie attributes as available for the multi-lists presentation of recommendations, as suggested by the literature [41, 31, 29, 11]. If more than one calculation was available for any given attribute, they were combined in a hybrid. Outside of these, one baseline was created that selected random movies. In total, eleven dif-

<sup>1</sup>MovieLens dataset location: <https://grouplens.org/datasets/movielens/latest/>

<sup>2</sup>TMDB: The movie database <https://developers.themoviedb.org/3>



Table 3.2: Table over the similarity functions that constitute recommendation lists selected from Table 3.3.

<b>Function</b>	<b>Description</b>
Baseline	Random unique pick from the 2512 movies in the dataset
All	Hybrid of all functions listed below
Title	Hybrid of the 5 different title similarity functions
Image	Hybrid of the 6 different image similarity functions
Plot	Hybrid of the 2 different plot similarity functions
Genre	Hybrid of the 2 different genre similarity functions
Director	Jaccard-based similarity function on director
Date	Release date distance based similarity function
Stars	Jaccard-based similarity function on actors
SVD	SVD-based similarity on tags
Tags	Tag metadata cosine

ferent similarity calculation functions were selected. An overview of the selected calculations can be seen in Table 3.2.

Table 3.3: Overall pool of available similarity functions, taken with permission Trattner &amp; Jannach study [54].

Name	Metric	Elaboration
Title:JW	$sim(r_i, r_j) = 1 -  dist_{JW}(r_i, r_j) $	Title Jaro-Winkler Distance based similarity
Title:LV	$sim(r_i, r_j) = 1 -  dist_{LEV}(r_i, r_j) $	Title Levenshtein Distance based similarity
Title:LCS	$sim(r_i, r_j) = 1 -  dist_{LCS}(r_i, r_j) $	Title Least Common Subsequence Distance- based similarity
Title:BI	$sim(r_i, r_j) = 1 -  dist_{BI}(r_i, r_j) $	Title Bi-Gram Distance- based similarity
Title:LDA	$sim(r_i, r_j) = \frac{LDA(Title(r_i)) \cdot LDA(Title(r_j))}{\ LDA(Title(r_i))\  \ LDA(Title(r_j))\ }$	Title LDA Cosine-based similarity
Image:BR	$sim(r_i, r_j) = 1 -  BR(r_i) - BR(r_j) $	Image Brightness Distance- based similarity
Image:SH	$sim(r_i, r_j) = 1 -  SH(r_i) - SH(r_j) $	Image Sharpness Distance- based similarity
Image:CO	$sim(r_i, r_j) = 1 -  CO(r_i) - CO(r_j) $	Image Contrast Distance- based similarity
Image:COL	$sim(r_i, r_j) = 1 -  COL(r_i) - COL(r_j) $	Image Colorfulness Distance- based similarity
Image:EN	$sim(r_i, r_j) = 1 -  EN(r_i) - EN(r_j) $	Image Entropy Distance- based similarity
Image:EMB	$sim(r_i, r_j) = \frac{EMB(r_i) \cdot EMB(r_j)}{\ EMB(r_i)\  \ EMB(r_j)\ }$	Image Embedding Cosine- based similarity
Plot:LDA	$sim(r_i, r_j) = \frac{LDA(Plot(r_i)) \cdot LDA(Plot(r_j))}{\ LDA(Plot(r_i))\  \ LDA(Plot(r_j))\ }$	Plot LDA Cosine- based similarity (LDA = LDA vector)
Plot:COS	$sim(r_i, r_j) = \frac{TFIDF(Plot(r_i)) \cdot TFIDF(Plot(r_j))}{\ TFIDF(Plot(r_i))\  \ TFIDF(Plot(r_j))\ }$	Plot Text Cosine- based similarity (TFIDF = TF-IDF weighted vector)
Genre:JACC	$sim(r_i, r_j) = \frac{\{Gen(r_i)\} \cap \{Gen(r_j)\}}{\{Gen(r_i)\} \cup \{Gen(r_j)\}}$	Genre Jaccard- based similarity
Genre:LDA	$sim(r_i, r_j) = \frac{LDA(Gen(r_i)) \cdot LDA(Gen(r_j))}{\ LDA(Gen(r_i))\  \ LDA(Gen(r_j))\ }$	Genre LDA Cosine- based similarity (LDA = LDA vector)
Dir:JACC	$sim(r_i, r_j) = \frac{\{Dir(r_i)\} \cap \{Dir(r_j)\}}{\{Dir(r_i)\} \cup \{Dir(r_j)\}}$	Director(s) Jaccard- based similarity
Date:MD	$sim(r_i, r_j) = 1 -  dist_{days}(r_i, r_j) $	Release Date distance- based similarity (unit = days)
Act:JACC	$sim(r_i, r_j) = \frac{\{Act(r_i)\} \cap \{Act(r_j)\}}{\{Act(r_i)\} \cup \{Act(r_j)\}}$	Actors Jaccard- based similarity
SVD	$sim(r_i, r_j) = svd(r_i, r_j)$	SVD-based similarity based on ratings Tag
Tags	$sim(r_i, r_j) = \frac{Tag(r_i) \cdot Tag(r_j)}{\ Tag(r_i)\  \ Tag(r_j)\ }$	Tag Genome Cosine- based similarity

### 3.1.3 Technical Details

While the framework is developed from the ground up for this thesis, several aspects are inspired by the preliminary study work of Trattner & Jannach. [54]. These aspects include the visual design of the interface and the structure of the study. With these elements pre-defined, the framework was developed to focus on the controllable multi-lists presentation of similar recommendations. The prototype is a web application developed using the Vue.js Javascript framework, with MongoDB Atlas handling data storage. The framework was developed as a single-page application, which entails that participants could only progress forward throughout the study. This design decision was made to disallow participants in the study to return to previous study steps and redo answers to make the study flow more similar to a scenario of regular browsing.

### 3.1.4 Recommendation Interface

This description of the prototype is divided into two parts. First, details of the multi-lists presentation of recommendations based on the previous section's similarity functions are discussed. Second, the details of the interaction elements for controllability are elaborated. The list sort methods concerning system interpretation of control are elaborated in the research design section, under conditions in Section 3.2.1. The controllable multi-lists recommender system can be seen in Figure 3.1.

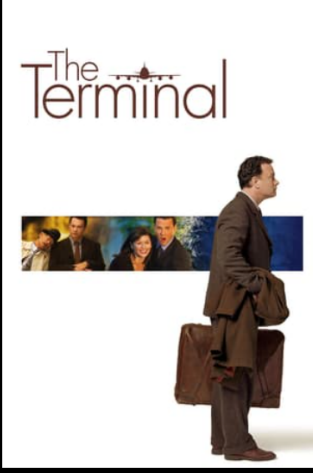
**Multi-lists presentation.** As a multi-lists presentation of recommendations, each similarity function constitutes a top-N list of movies. As list length is not considered to be substantially important [7] and to reduce the cognitive load for the participants when evaluating all these lists [31, 49], as well as noting standard practices in other studies[5, 41, 29], each list only contains five of the movies with the highest similarity score in context to the reference movie selected. The same movie may show up in multiple lists, which is also typical of commercial services such as Netflix and Amazon, and shown to be beneficial in a Zhao et al. study [58]. Each list is also labeled and displayed above each row, identical to the function name given in Table 3.2 with two exceptions. The "baseline" label is replaced with "Miscellaneous" when presented to the participants to avoid possible bias [23, 15]. The second exception was based on observations made during the limited test trials in which participants struggled to under-

stand what the label "SVD" entailed. As such, this was labeled as "Community Preference".

**Your Reference Movie:**

## The Terminal

**(2004)**  
 Viktor Navorski is a man without a country; his plane took off just as a coup d'etat exploded in his homeland, leaving it in shambles, and now he's stranded at Kennedy Airport, where he's holding a passport that nobody recognizes. While quarantined in the transit lounge until authorities can figure out what to do with him, Viktor simply goes on living – and courts romance with a beautiful flight attendant.  
 Stars: Tom Hanks, Catherine Zeta-Jones, Stanley Tucci, Chi McBride  
 Director: Steven Spielberg  
 Genres: Comedy, Drama



Please report on a scale from very unsatisfied (1) to very satisfied (5) on how you happy you are with the current list order, from top to bottom.






Very Unsatisfied      Very Satisfied  
 1      2      3      4      5

●      ●      ●      ●      ●

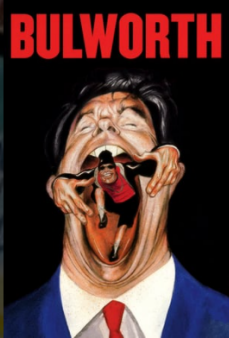

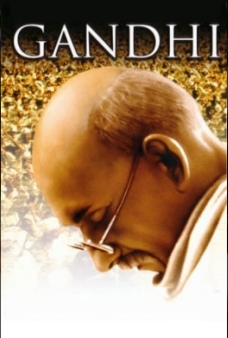
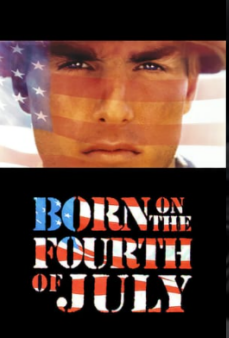

[Zoom Out](#)

Here are your movie recommendations, based on...

**Miscellaneous**

**Plot**

Community Preferences

Figure 3.1: The controllable multi-lists presentation of similar recommendations. The two interaction elements for control can be seen right of lists. More recommendation lists can be viewed by scrolling.

**Interaction for control.** The interaction elements that enable users to enforce their control on the system are the arrows seen in Figure 3.1. As guided by RQ1, the intended goal is to allow users control over feature importance, which is implemented by having a multi-list presentation of recommendations in which each list is based on a unique feature or method. Based on an investigation into literature, no studies were found investigating control and user interaction in such a context. Therefore, the selection is based on the theoretical foundation and suggestions made by literature review on similar topics [31, 11, 52, 51, 27, 39, 45] to provide more grounded assumptions on selected controls utility and ease of handling in this context.

To this end, two interaction method was selected for user enforcement of control. The first interaction method included was drag & drop, in which users can drag a list to any position by a handle, which is the double-sided arrow seen in Figure 3.1. This interaction method was selected based on an assumption of utility in this context and inspired by how it was utilized in the *uRank* system by C. Di Sciascio et al. [16]. A second alternative was also incorporated, which constitutes simple buttons shaped like arrows, which moves a list of one index in the arrow's direction, which is noted to have been used in some similar fashion in a Harper et al. study [25].

## 3.2 Research Design

This section details the research design for evaluating the prototype following the research questions. It begins by providing a general summary of the conditional quantitative study before detailing the conditions selected and further describing each step of the study. Following this, the metrics of this study are described, and this section ends with an overview of the statistical analysis methods.

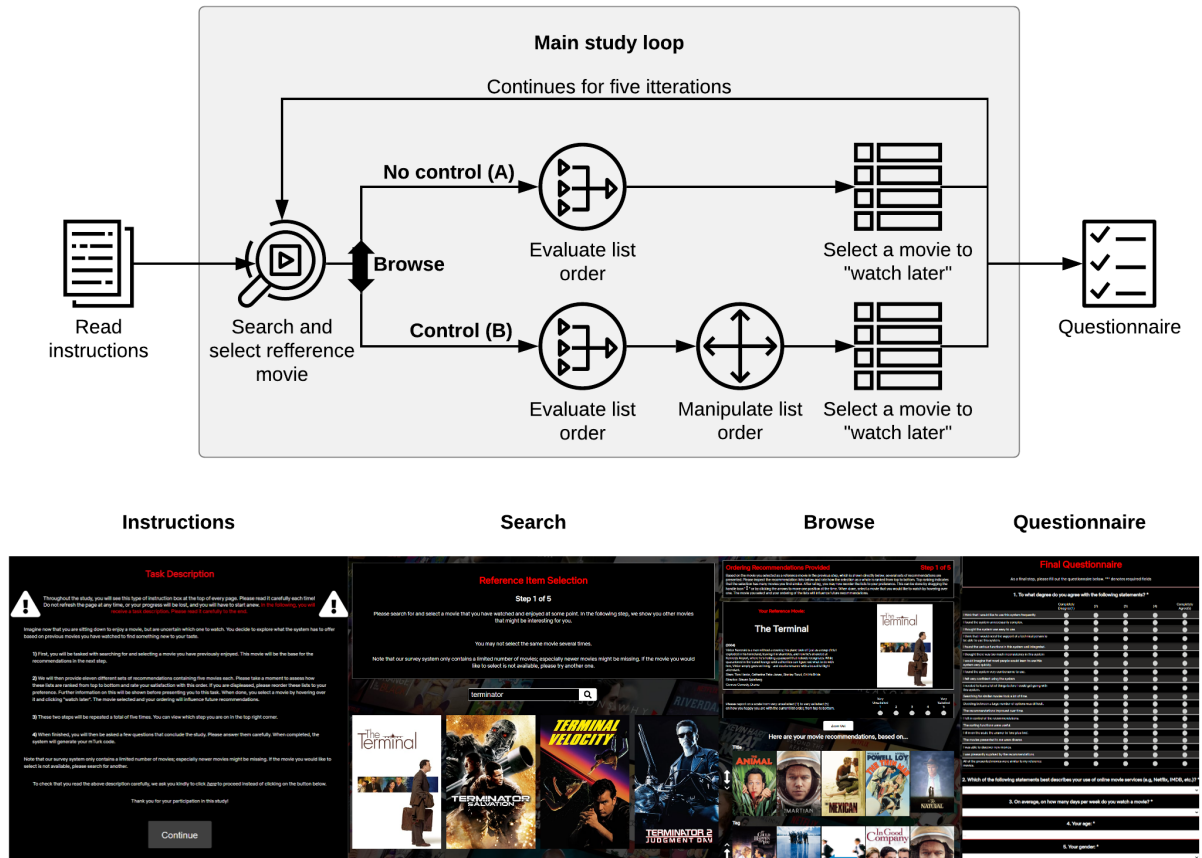


Figure 3.2: General overview of the study, with associated screenshots of the study for each step. Full views of these screens are available in the Appendix A.

The prototype described forms the core of the study and is expanded to include relevant functionality. A visual flowchart of the study can be seen in Figure 3.2. At the start, the participant is presented to the study and given a complete list of instructions. After confirming, the primary study loop that lasts for five iterations begin. Here, participants search and select a reference movie and are then given a multi-lists presentation of recommendations. How the system orders the list, and if control is enabled are conditional dependent and further detailed in Section 3.2.1, but in any condition, the participant first evaluates the list order. This browsing ends when participants one of the recommended movies to watch later in which the user is tasked with searching for a new reference movie. After five iterations, the participants are given a questionnaire and tasked with evaluating the system. When this is handed in, the study is complete.



### 3.2.1 Conditions

The study method is a quantitative conditional evaluation to find variances with pairwise comparison to answer research questions. With this in mind, the study is divided into several conditions depending on the research questions. For evaluating controllability in the multi-lists presentation of recommendation (RQ1), two conditions are needed in terms of the prototype with and without controllability. For the effect of different list sort methods (RQ3), three methods were selected as sub-conditions, meaning that the study totals six conditions. An overview of all conditions can be seen in Table 3.4

Table 3.4: Overview of conditions and names.

		Conditions on control	
		No control (A)	Control (B)
Sub-conditions on list sort method	Random sort (1)	No control random sort (A-1)	Control random sort (B-1)
	Fixed sort (2)	No control fixed (A-2)	Control fixed sort (B-2)
	Weighted sort (3)	No control weight sort (A-3)	Control weight sort (B-3)

The following is an elaboration of the list sort methods chosen for this study. Pragmatically, the list sort methods are just different methods for the system to reorder the hierarchical order of the lists in a multi-list presentation based on the user control.

The first list ordering chosen method is a random Fisher-Yates shuffle [17] that is performed per iteration step. This means that the user control has no effect on the system and is chosen as a baseline comparison. This shuffling method is also employed in the other sub-conditions' first iteration to provide an unbiased starting point [23].

The second method chosen is fixed sort. In this method, after the first shuffle, no further sorting is done by the system. If control is enabled, the order provided by participants is preserved between iterations. This interaction can be described as a Markov Chain as the only dependent state is the previous state [38].

Weight sort is the final method. This method, which has been utilized in different forms in discussed studies [44, 36, 8] takes the way of applying weights to list based on index positions of each list per iteration. The weight for any given list is calculated based on this formula:  $w_l = \sum_{i=1}^z \frac{1 + y - x}{y}$  where  $z$  is the iteration,  $y$  is the total number of lists present in the system, and  $x$  is the index value of the list counting from the top. An exception to this algorithm is

that the list from which a participant selects a movie is always considered the top position.

As a reference guide based on Table 3.4, the conditions on controllability is referred to as follows: no control (A) and control (B). The list sort conditions consist of data from both controllability conditions and are referred to with integers with (1) representing random sort, (2) fixed sort, and (3) weighted sort. Specific conditions may be referred to based on controllability and list sort methods, such as no control (A-3) for an interface without control with weighted sort or control (B-1) for an interface with control and random sort.

### 3.2.2 Study Details

Mturk<sup>3</sup>, Amazon's crowdsourcing platform, was opted for as the tool to collect the quantitative data needed through a user study. This platform has been utilized in the preliminary study by Trattner & Jannach [54]. Additionally, a previous study by Hauser & Schwarz [26] on the quality of participants' work supports its use in this study [26]. As the focus of this study is to evaluate the viability of a controllable multi-list presentation of recommendations, participants could only partake in the study on a personal computer. This limitation is put into place to have a homogenous testing environment and prevent different platforms with other interactive elements such as touchscreens and smaller visual displays from distorting results [23, 52]. Participants had to have a Mturk score above an arbitrary threshold to reduce potential inattentiveness [54, 26]. The average completion time was 12.6 minutes, in which participants have been compensated one dollar for their efforts.

The study structure is based on the preliminary study [54] with a key difference in that several iterations are included. The number of iterations settled on was five, which is both assumed to be enough to measure some results while still short enough to be both economical and practical. The number of iterations required by participants is explained in the introduction, and the current iteration is always displayed to the participants.

---

<sup>3</sup>Mturk platform: <https://www.Mturk.com>



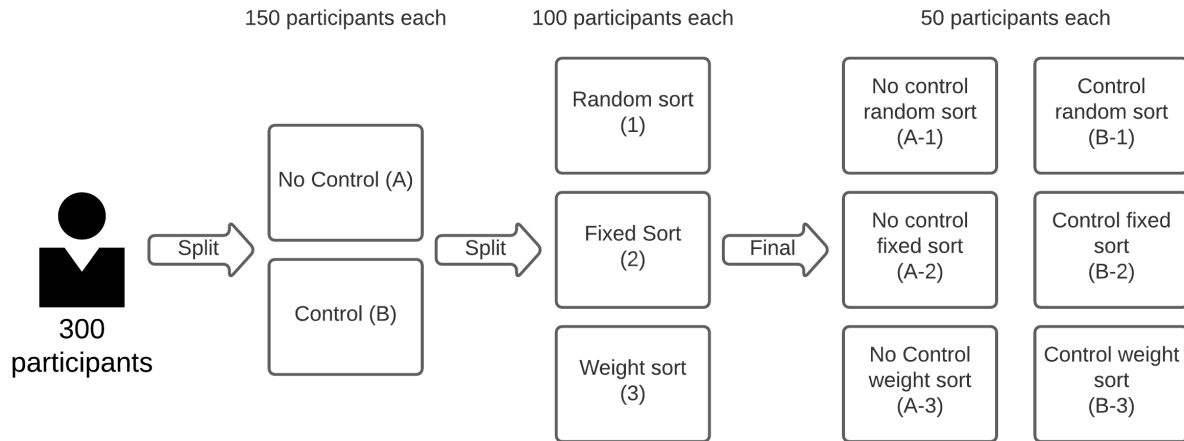


Figure 3.3: Overview of the sample distribution per condition of this study.

The study was performed in batches as visualized in Figure 3.3. Each batch consisted of 50 participants focusing on one condition. A total of 300 participants was recruited for this study. To ensure the independence of samples, participants from previous batches are excluded from future participation. Due to some technical errors, some entries had to be omitted due to incomplete data. This consisted of three participants for A-1, two for A-2, and one for B-2. A supplement run to fill these data gaps was performed by recruitment among the student populace with no prior knowledge or engagement with the project. As the supplement sample size is small compared to the total sample size, any bias induced by local recruitment is considered negligible.

### 3.2.3 Study Phases

The following paragraphs detail each of the four main phases of the study; instructions, search phase, browsing phase, and survey phase.

**Instruction Phase.** The first step of the study after a participant is redirected from Mturk is the introduction screen. On this page, the participants are presented with the surrounding context and general task, an overview of the study, and their general instructions. These instructions are identical across conditions except that elaboration on control is omitted from the no control (A) conditions. Additionally, the participants are informed that their inputs will influence future recommendations without any further elaboration, influenced by Harper et al. study [25].

Participants then move on by confirming that they have read the instructions. An attention check for filtering purposes was implemented here by writing instructions to click on a button hidden in the text rather than the large button at the bottom, but the results found only 27% of all participants passed this test and was therefore dropped for being too strict. Screenshot of the instruction screen can be seen in the Appendix A as Figure A.1.

**Search Phase.** When a participant navigates from the instruction screen, the main study loop begins, a two-step process that lasts for five iterations. The first step is the search phase. In this phase, participants are tasked with searching for a reference movie for which the system will provide recommendations. While time demanding, this has been shown in the Mc-Nee et al. study [39] to increase user satisfaction. Participants are not allowed to select a reference movie more than once. This page is visualized in the Appendix A under Figure A.2.

**Browsing Phase.** When a reference movie is selected, the next step is the browsing phase. Here, the participants are presented with a multi-lists presentation of recommendations based on the selected reference movie. Instructions are reiterated and elaborated on top of the screen, following an information screen that displays the meta-information of the selected reference movie. Under these, the eleven rows that constitute the multi-list presentation of recommendations are displayed. Based on feedback from the trial, a zoom button that shrinks the recommendations view by 50% is implemented to provide participants with an optional overview of all recommendations.

The first task given to the participants here is to rate the order of the lists on a five-point Likert scale from "very unsatisfied" to "very satisfied". Participants may not perform any control action or continue from this phase before answering this question. If attempted, a visual error trigger. Following this, participants in the control (B) conditions are tasked with reordering the lists from the top (positive) to the bottom (negative). When they are satisfied with the order, they select a movie from one of the lists to "watch later". The no control (A) participants do the same after rating the order. This page is visualized in the Appendix A under Figure A.4.

One additional element to the browsing phase included for the control (B) participants in the first iteration only is one more instruction screen. This screen provides a tutorial of how participants can use control to rearrange the list order by showcasing the control elements using .gifs and describing possible actions. This screen can be seen in Appendix A as Figure A.3.

**Survey Phase.** Participants move between the search phrase and the browsing phase. When five iterations have passed, the participants are redirected to an end of study questionnaire. In this survey phase, participants are tasked with evaluating the study by answering questions, further described in Section 3.2.4. When done, the study ends.

To summarize and provide an overview of the study, detailed process charts that elaborate on the study structure per condition are detailed in Figure 3.4 for the no control conditions (A) and Figure 3.5 for the control conditions (B).

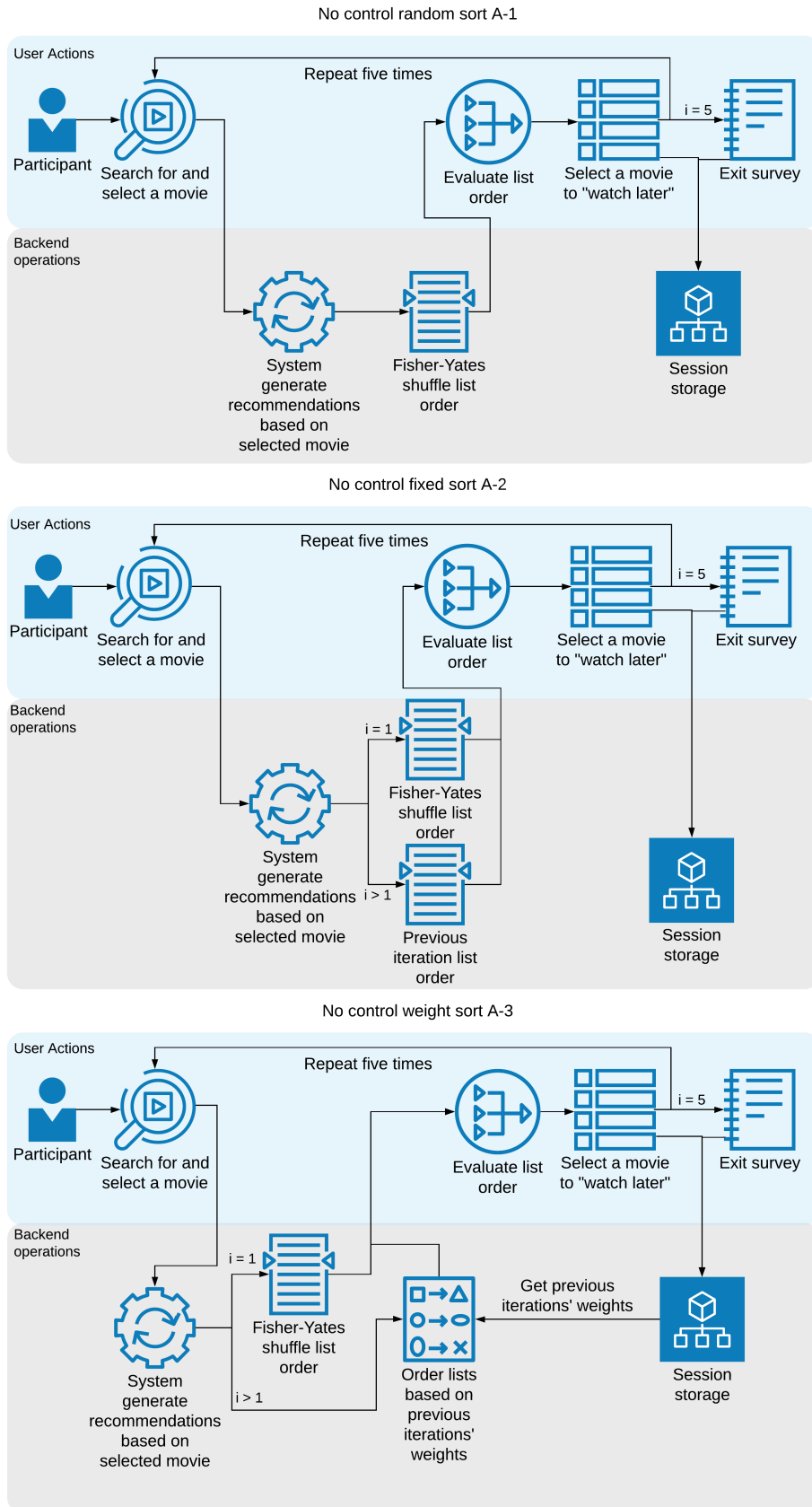


Figure 3.4: Process chart highlighting the operations of no control (A).

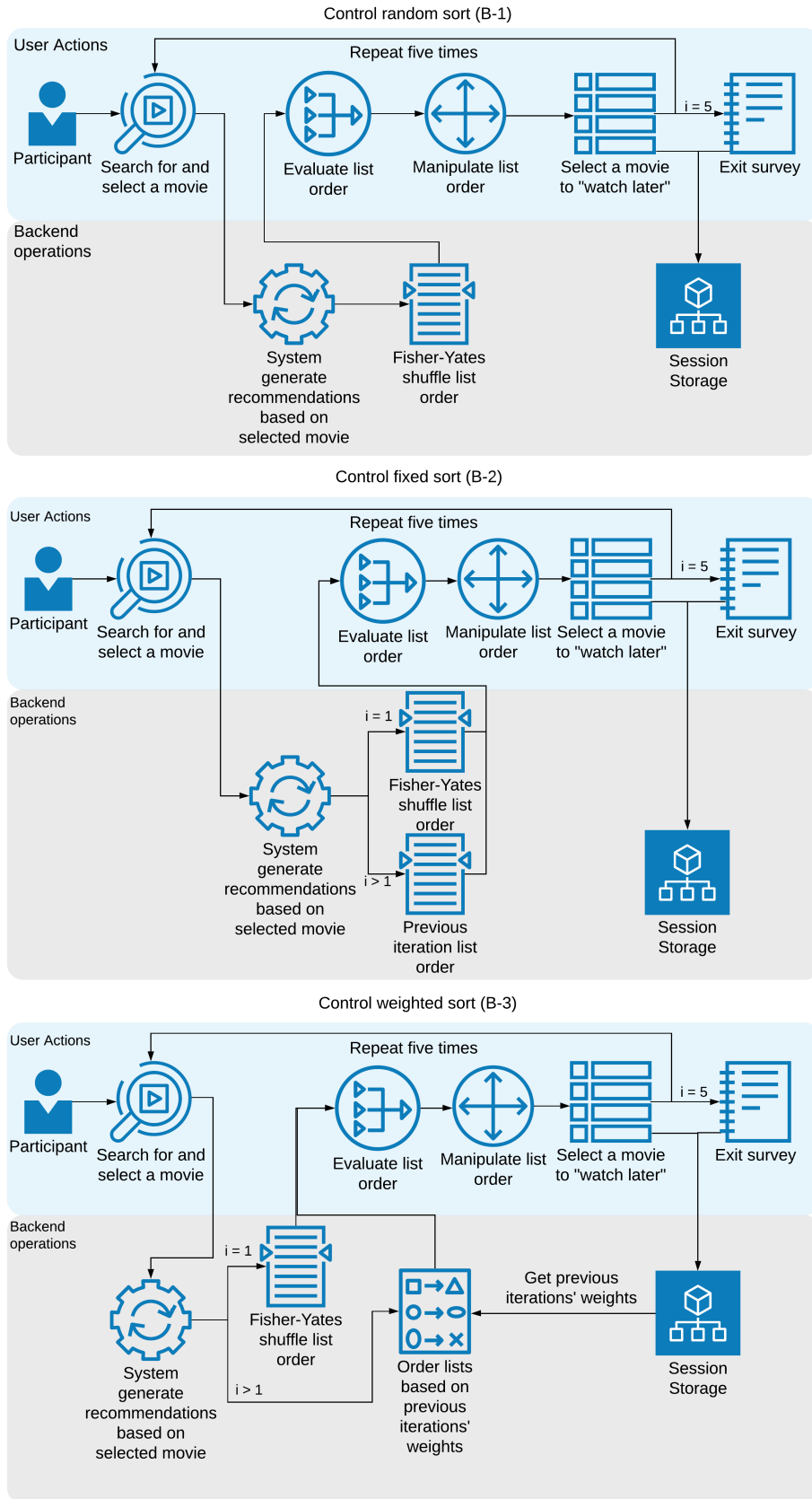


Figure 3.5: Process chart highlighting the operations of control (B).

### 3.2.4 Metrics

The metrics selected for the study are detailed in the following, which includes the tracked activity of participants and the evaluation questions given. These questions are organized into categories [33] based on evaluating certain aspects of the system and are selected based on the preliminary study [54] and similar research. A complete list of the questions can be seen in Table 3.5. The elaboration begins by describing each question category in relation to the associated research questions.

**System usability.** This category consists of the commonly used system usability scale [37, 23], which is selected for this study as a holistic evaluation of the system. This is used for RQ1, RQ2, and RQ4. The system usability scale is made up of ten Likert-scale and a final score based on these [37]. The principal metric here is the overall score, but the questions themselves are included in the statistical analysis for a higher granular view of the results.

**Recommendation quality.** Participants' subjective view of the recommendations provided may differ depending on conditional properties, as noted in Pu et al. study [29]. For this, five metrics are selected based on important evaluation aspects of the recommender system, such as serendipity, relevancy, diversity, and novelty [54, 57, 25]. These are of interest to RQ1, RQ2, and RQ4. In addition, list order satisfaction is included as a measurement of participant satisfaction with the current list ordering per iteration. This metric is the primary metric to gauge the effect of list sorting methods and are included for RQ3.

**System satisfaction.** This category consists of questions selected to evaluate specific parts of the implemented controls, such as the usefulness of sorting functions, recommendation improvement over time, and participants' evaluation of the usefulness of control. These are of interest to RQ1, RQ2, and RQ4.

**Demographic** Similar works have shown that demographic properties users may influence user evaluation of recommender system [21, 6, 40, 48, 47, 6, 35]. These are included for RQ4 to discover if these factors are influential when evaluating controllability. These are the age and gender of the participants, experience with similar systems, and domain knowledge.

Table 3.5: Overview of the questions posed in the questionnaire at the end of the study, organized by metric group. Includes metric name, possible values and exact question posed.

Metric	Options	Question
<i>System Usability</i>		
SUS Score	-	Calculated score based on questions the below [37] .
Wish to use	Likert-Scale 1-5	I think that I would like to use this system frequently.
Complexity	Likert-Scale 1-5	I found the system unnecessarily complex.
Easy to use	Likert-Scale 1-5	I thought the system was easy to use.
Assistance need	Likert-Scale 1-5	I think that I would need the support of a technical person to be able to use this system.
Functions integration	Likert-Scale 1-5	I found the various functions in this system well integrated.
Inconsistencies	Likert-Scale 1-5	I thought there was too much inconsistency in this system.
Easy to learn	Likert-Scale 1-5	I would imagine that most people would learn to use this system very quickly.
Cumbersome	Likert-Scale 1-5	I found the system very cumbersome to use.
Confidence using	Likert-Scale 1-5	I felt very confident using the system.
A lot to learn	Likert-Scale 1-5	I needed to learn a lot of things before I could get going with this system.
<i>Recommendation Quality</i>		
List order satisfaction N**	Likert-Scale 1-5	Please report on a scale from very unsatisfied (1) to very satisfied (5) on how you happy you are with the current list order, from top to bottom.
List order satisfaction**	Avg	Mean value of list order satisfaction after five iterations
Diversity	Likert-Scale 1-5	The movies presented to me were diverse.
Novelty	Likert-Scale 1-5	I was able to discover new movies.
Serendipity	Likert-Scale 1-5	I was pleasantly suprised by the recommendations.
Relevance	Likert-Scale 1-5	All of the presented movies were similar to my reference movies.
<i>System Satisfaction</i>		
Recommendation sim	Likert-Scale 1-5	Searching for similar movies took a lot of time.
Difficult selecting movie	Likert-Scale 1-5	Deciding between a large number of options was difficult.
Recommendation Improved	Likert-Scale 1-5	The recommendations improved over time.
Felt in control	Likert-Scale 1-5	I felt in control of the recommendations.
Sorting function useful	Likert-Scale 1-5	The sorting functions were useful.
<i>Demographic</i>		
Streaming service usage	Daily(1)-Once a week/Once a month(2)*-Once every three months(3)-Hardly(4)	Which of the following statements best describes your use of online movie services (e.g, Netflix, IMDB, etc.)?
Movies per week	Option range: 0 to 7	On average, on how many days per week do you watch a movie?
Age category	Between 18 and 100	Your age
Gender	Female(1) - Male(2) - Other (3)	Your gender

\*The option of "Once a month" was available in the study but due to a technical error the value was registered as once a week.

\*\* List order satisfaction are not asked during the survey phase, but during each browse phase.

Outside of these four main categories of questions, tracked activity by participants are also a group of metrics. These metrics are the logged activity of participants. For control elements, this is the number of interactions performed using the control elements and position of lists selected, as it is of interest to evaluate the two interaction methods chosen for control in RQ2. These logged activity-based metrics are summarized in table 3.6

Table 3.6: Overview of the tracked activity metrics used in this study.

Metric	Value	Explanation
Total list ordering	Avg value	The distance a list have been moved in terms of absolute value of index change.
Click list ordering	Avg value	The distance a list have been moved using clicks in terms of absolute value of index change.
Drag&Drop list ordering	Avg value	The distance a list have been moved using drag&drop in terms of absolute value of index change.
Times click was used	Avg value	Number of times clicks was utilized
Times drag&drop was used	Avg value	Number of time a drag&drop operation was performed
Movie index in list	Avg value	When participants select a movie, it's index in a list is stored
Selected list index	Avg value	When participants select a movie, the index of the list is stored
Preferred Control Element	0 = No preference, 1 = Drag & Drop 2 = Click	Categorical variable based on which interaction method a participant utilized the most, based on distance of each control.

Based on the background research, the metrics were selected primarily due to their relevance in evaluating the research questions and on other noted practices [23]. A single sample consisting of one participant will tally a total of 140 metrics based on tracked activity and data per iterations, with only the ones relevant to the research questions included in this study. As such, the data gathered are available for future research. For more information on this, Section 5.3 detail unexplored regions of the dataset and Section 5.4 details information on accessing this dataset.

### 3.2.5 Statistical Analysis

The method for statistical analysis is to look at statistically significant variances between conditions on associated metrics. To this end, two main statistical methods are chosen. The student t-test is utilized for pairwise comparison between conditions, with Cohen's  $d$  to calculate the effect. For a comparison between several groups, the one-way analysis of variance (ANOVA) is utilized with omega squared ( $\omega^2$ ) as the effect metric. For the ANOVA test that returns a statistically significant value ( $p < 0.05$ ), the Tukey-HSD post hoc test is used with Cohen's  $d$  for effect. The following is an organized description of how each analysis was performed in relation to research questions.



**RQ1: Evaluating controllability in multi-lists.** The principle analysis is t-tests is performed between no control(A) and control (B) conditions on the discussed metrics in Section 3.2.4. These pairwise comparison analyses are also performed between sub-conditions to isolate list-sort methods. Finally, an ANOVA test between list-sort methods is performed on the metric groups to detect if an interaction effect is present between control and list sort conditions.

**RQ2: Evaluating selected interaction methods.** To analyze if the selected interaction methods impact participant evaluation of the system, the participants are grouped by which interaction method they ordered the lists with the most. A pairwise comparison using the t-test is then performed between click and drag & drop users on the selected metric groups.

**RQ3: Evaluating list-sort methods.** The primary metric to analyze is the list order satisfaction reported by participants, both per iteration and the average value. A one-way ANOVA achieves this by looking for variances between the three list sort conditions. The analysis was performed on both aggregated and separated control condition data. The one-way ANOVA on interaction effect performed for RQ1 is also relevant here to look at possible variances in other metrics.

**RQ4. Evaluating demographic influences.** The final set of statistical analyses are a comparison analysis between groups based on demographic metrics. This begins with including demographic variables in all previous analyses to verify distribution among conditions. A comparative analysis is performed on the evaluation metrics to see if any demographic properties impact the system's evaluation. A t-test is utilized if a demographic metric consists of only two categories, one-way ANOVA if several are present.

### 3.2.6 Data Filtering

A total of 300 participants conducted the research tasks (Figure 3.3). Based on common practice [26, 54], participants who failed an attention check are omitted from the data analysis. Initially, it was planned to utilize an attention check presented at the beginning of the study to filter it. However, only it was considered too conservative with only 25% passing. Due to

Table 3.7: Sample size overview for the conditions.

List sort conditions		No control conditions		Control conditions	
Total	Filtered	Total	Filtered	Total	Filtered
$N_1 = 100$	$N_1 = 94$	$N_{A-1} = 50$	$N_{A-1} = 47$	$N_{B-1} = 50$	$N_{B-1} = 47$
$N_2 = 100$	$N_2 = 88$	$N_{A-2} = 50$	$N_{A-2} = 46$	$N_{B-2} = 50$	$N_{B-2} = 42$
$N_3 = 100$	$N_3 = 80$	$N_{A-3} = 50$	$N_{A-3} = 52$	$N_{B-3} = 50$	$N_{B-3} = 38$

this issue, the selection criteria were modified only to utilize the secondary attention check, where an arithmetic question was present in the survey. Based on this filtering, the remaining samples per condition used for the statistical analysis can be viewed in Table 3.7.





# Chapter 4

## Results

This chapter details the results from the statistical analysis performed on the study data. It is organized by which research question the analysis pertain to as follows:

- **RQ1:** Section 4.1 details the analysis results performed to evaluate controllability in multi-lists presentation of recommendations.
- **RQ2:** Section 4.2 details analysis results on how evaluation differed between interaction methods.
- **RQ3:** Section 4.3 details analysis results on how list sort methods affected the participants' list order satisfaction.
- **RQ4:** Section 4.4 details the analysis performed to evaluate if demographic aspects of the participants influenced their evaluation of the system.

## 4.1 Evaluation of Control Elements (RQ1)

Pairwise comparisons between the no control (A) and control (B) groups were performed between the conditions as a whole and organized by sub-conditions on list-sort methods. The results showed persistent trends across all t-tests, yet the number of statistically significant metrics decreased as list-sort method complexity increased. A one way ANOVA was performed on list sort methods as a whole and list-sort sub-conditions under control conditions to discern if this diminishing effect signified an interaction effect. The ANOVA test did not find any significant variances, so no interaction effect is assumed to take place even if weighted sort (Table B.3) had barely any significant statistical variances as opposed to random sort (Table B.1). The sub-conditional t-tests are available in Appendix B.1.2 with the ANOVA for interaction effect in Appendix B.1.3. Following the table is an elaboration of results per metric category.

Table 4.1: Pairwise comparison with t-test between no control and control overall. Means are displayed with standard error and are underlined in metrics where objective better values are present. N denotes sample size. Note: † = significant metric, \* $p < .05$ , \*\* $p < .01$ , \*\*\* $p < .001$ .

	$N_A = 135$		$N_B = 127$	
<i>System Usability</i>	Mean <sub>A</sub>	Mean <sub>B</sub>	$p$	Cohen's $d$
SUS Score†	<u>73.7±1.60</u>	65.4±1.76	>.001***	.43
Wish to use	<u>3.64±0.10</u>	3.50±0.11	.305	.13
Complexity†	<u>2.1±0.11</u>	2.59±0.11	.002**	-.40
Easy to use†	<u>4.18±0.09</u>	3.89±0.09	.019*	.29
Assistance need†	<u>1.71±0.11</u>	2.06±0.12	.031*	-.27
Functions integration	<u>3.72±0.08</u>	3.54±0.10	.163	.17
Inconsistencies†	<u>2.39±0.11</u>	2.73±0.12	.039*	-.26
Easy to learn†	<u>4.18±0.08</u>	3.79±0.08	>.001***	.41
Cumbersome†	<u>2.12±0.11</u>	2.74±0.12	>.001***	-.48
Confidence using	<u>4.01±0.08</u>	3.89±0.09	.349	.12
A lot to learn†	<u>1.92±0.11</u>	2.33±0.12	.010*	-.32
<i>Participant Activity</i>				
Movie index in list†	1.86±0.07	1.55±0.07	>.001***	.41
Selected list index†	2.15±0.14	1.50±0.18	.004**	.35
<i>Recommendation Quality</i>				
List order satisfaction	<u>3.93±0.07</u>	3.86±0.07	.450	.09
Diversity†	<u>3.94±0.08</u>	3.70±0.08	.034*	.26
Novelty	<u>3.88±0.09</u>	3.74±0.10	.292	.13
Serendipity	<u>3.72±0.09</u>	3.52±0.09	.124	.19
Relevance	<u>3.37±0.10</u>	3.30±0.10	.626	.06
<i>System Satisfaction</i>				
Recommendation sim†	2.27±0.11	<u>2.69±0.12</u>	.009**	-.32
Difficult selecting movie	<u>2.65±0.11</u>	2.79±0.11	.400	-.10
Recommendation Improved	3.26±0.10	<u>3.38±0.10</u>	.395	-.11
Felt in control	<u>3.41±0.11</u>	3.24±0.10	.263	.14
Sorting function useful	3.47±0.10	<u>3.54±0.10</u>	.625	.06
<i>Demographic</i>				
Age category	2.81±0.09	2.80±0.09	.927	.01
Streaming service usage	1.67±0.06	1.69±0.06	.831	-.03
Gender	1.53±0.04	1.57±0.05	.603	-.06
Movies per week	3.10±0.17	2.91±0.16	.433	.10

**System evaluation.** In terms of system evaluation, Table 4.1 shows a statistical significant negative variance in SUS score when control is implemented ( $p < 0.001$ ,  $d = .43$ ). Through a further investigation into the questions that form the overall system evaluation score, it is made clear that no control was more positively evaluated, with 7/10 questions posed as statistically significant. These differences become less significant when sub-conditions are paired based on list-sort methods, but at no point is control (B) evaluated more positively than no control (A). Control therefore appears to have an adverse effect on system usability.

**Recommendation quality.** Participants found the no control (A) recommendations more diverse ( $p = .048$ ,  $d = .63$ ) with no other metrics statistically significant. Surprisingly, list order satisfaction did not improve in control (B). In short, control does not appear to affect participant evaluation of recommendation quality.

**System satisfaction.** While there are no significant variances found in terms of participants reporting that recommendations were improving, felt more in control, or that the sorting functions were useful, control was evaluated to provide more similar recommendations ( $p = 0.009$ ,  $d -0.32$ ). It is also noted that in the pairwise comparison with weighted sort as seen in Table B.3, participants also reported that the recommendations improved much more than no control participants ( $p = 0.038$ ,  $d -0.47$ ). To summarize, control appears to affect some aspects of system satisfaction positively.

**Summary.** Compared to the no control interface, control was negatively evaluated by participants in terms of overall system evaluation, appeared to have a negligible effect on recommendation quality, and some positive effects on some aspects of system satisfaction.

**Other results.** It is noted that participants in the no control (A) condition had a tendency to select movies both deeper further down in the multi-list presentation ( $p < .001$ ,  $d = -0.41$ ) and further down the list ( $p = .004$ ,  $d -0.35$ ).



## 4.2 Evaluation Variances by Interaction Preference (RQ2)

Subsequently, a pairwise comparison between drag & drop(DD) and click(C) was. The comparison is on participants' utilization of each interaction method, based on the assumption that the method with the highest move distance by a user indicates a preference. For this purpose, all participants who did not interact with the system are omitted from the analysis. There were no noted occurrences where distance was equal between methods for any participant. The data for this analysis is based on the entirety of control(B). The pairwise t-test results can be viewed in Table 4.2.

The first aspect noted is that among the 150 participants in the no control(B), 51 never interacted with the controls; 76 mostly utilized the buttons while only 23 preferred drag & drop. This low sample size of drag & drop users may impact the reliability of the results.

Based on the mean values and statistical significance noted, drag & drop users occasionally interacted with the arrow buttons with an average interaction count of 2.47(DD) versus 7.6 (C), yet click users would barely use the drag & drop functionality with 3.16(DD) versus 0.29(C). While differences are expected when grouping users based on which interaction method they primarily used, the low value of the alternative interaction method signifies a tendency to pick one method and stick with it. Drag & Drop users also moved lists more than click users, with an average index reorganization of 12.15(DD) versus 8.27(C).

While the differences in participant activity are significant, this does not seem to affect the overall system evaluation, recommendation quality, and system experience evaluation to any significant degree. These results may be due to the low sample size of DD preference users. However, it is noted that in terms of demographic, drag & drop users tended to be more experienced with similar systems than their counterparts.

Table 4.2: Pairwise comparison with t-test between utilization of control elements by participants. DD = Drag & Drop users, C = Click users. Means are displayed with standard error and are underlined in metrics where objective better values are present. N denotes sample size. Note: † = significant metric, \* $p < .05$ , \*\* $p < .01$ , \*\*\* $p < .001$ .

	$N_{DD} = 22$		$N_C = 70$		
<i>System Usability</i>	Mean <sub>DD</sub>	Mean	Mean <sub>C</sub>	$p$	Cohen's $d$
SUS Score	<u>70.0±4.92</u>		65.3±2.38	.357	-.23
Wish to use	3.18±0.29		<u>3.41±0.14</u>	.438	.19
Complexity	<u>2.14±0.27</u>		2.6±0.15	.144	.36
Easy to use	<u>4.09±0.22</u>		3.77±0.12	.190	-.32
Assistance need	<u>1.64±0.25</u>		1.83±0.13	.477	.17
Functions integration	3.41±0.23		<u>3.44±0.13</u>	.898	.03
Inconsistencies	2.77±0.34		<u>2.53±0.14</u>	.449	-.19
Easy to learn†	<u>4.14±0.20</u>		3.64±0.12	<b>.043*</b>	-.50
Cumbersome	<u>2.45±0.32</u>		2.69±0.15	.475	.18
Confidence using	<u>4.09±0.22</u>		3.77±0.13	.228	-.30
A lot to learn	<u>1.91±0.29</u>		2.27±0.15	.247	.28
<i>Participant Activity</i>					
Total list ordering†	12.15±1.39		8.27±0.77	<b>.016*</b>	-.6
Click list ordering†	2.47±0.67		7.60±0.68	<b>&gt;.001***</b>	.99
Drag&Drop list ordering†	9.67±1.11		0.67±0.23	<b>&gt;.001***</b>	-.30
Times click was used†	2.47±0.67		7.60±0.68	<b>&gt;.001***</b>	.99
Times drag&drop was used†	3.16±0.39		.29±0.10	<b>&gt;.001***</b>	-2.49
Movie index in list	1.50±0.13		1.51±0.09	.948	.02
Selected list index	1.55±0.38		1.09±0.21	.271	-.27
<i>Recommendation Quality</i>					
List order satisfaction	3.48±0.18		<u>3.81±0.09</u>	.090	.42
Diversity	<u>3.77±0.16</u>		3.64±0.11	.559	-.14
Novelty	<u>3.73±0.22</u>		3.64±0.14	.768	-.07
Serendipity	3.32±0.27		<u>3.43±0.12</u>	.675	.10
Relevance	<u>3.23±0.23</u>		3.13±0.13	.715	-.09
<i>System Satisfaction</i>					
Recommendation sim	2.41±0.33		<u>2.63±0.15</u>	.510	.16
Difficult selecting movie	<u>2.59±0.30</u>		2.69±0.15	.760	.07
Recommendation Improved	<u>3.27±0.27</u>		3.23±0.12	.870	-.04
Felt in control	3.09±0.23		<u>3.14±0.13</u>	.844	.05
Sorting function useful	<u>3.59±0.29</u>		3.43±0.13	.577	-.14
<i>Demographic</i>					
Age category	2.50±0.17		2.93±0.13	.086	.42
Streaming service usage†	1.32±0.10		1.77±0.10	<b>.014*</b>	.61
Gender	1.73±0.12		1.61±0.06	.362	-.22
Movies per week	3.09±0.43		2.79±0.21	.488	-.17

### 4.3 List Sort Methods Impact on System Satisfaction (RQ3)

The following analysis was performed to evaluate the effect of different system enforcement of control in list-sort methods (RQ3). Results from the interaction effect analysis of Section 4.1 exhibited no effect between control and list sort method on other metrics, which also pertain to list sort methods not affecting this selection of metrics.

With these initial results in mind, this statistical analysis focuses on the list order satisfaction given by participants each iteration and the mean value across all iterations, which was compared without considering control, and when considering control. Overview of participants' evaluation of list order can be seen in Figure 4.1 with mean values and standard error. There appear to be some noted differences in the figure, but a one-way ANOVA analysis found no significant variances as seen in Table 4.3. Consequently, the list sort method does not appear to have any effect on list order satisfaction as seen in this section, or any other metric explored as seen in Appendix B.5 and Appendix B.4.

Table 4.3: One-Way ANOVA on users' list order satisfaction score between conditions and sub-conditions. All metrics are rounded to nearest third decimal for p-value and second decimal for omega squared. It. denotes iteration number

	List Sort (1, 2, 3)		No Control (A-1, A-2, A-3)		Control (B-1, B-2, B-3)	
	N <sub>1</sub> = 94		N <sub>A-1</sub> = 47		N <sub>B-1</sub> = 47	
	N <sub>2</sub> = 88		N <sub>A-2</sub> = 46		N <sub>B-2</sub> = 42	
	N <sub>3</sub> = 80		N <sub>A-3</sub> = 52		N <sub>B-3</sub> = 38	
	<i>p</i>	$\omega^2$	<i>p</i>	$\omega^2$	<i>p</i>	$\omega^2$
List order satisfaction it.1	.930	-.01	.341	.00	.410	-.00
List order satisfaction it.2	.963	-.01	.431	-.01	.347	.00
List order satisfaction it.3	.306	.00	.791	-.01	.287	.00
List order satisfaction it.4	.760	.00	.331	-.00	.807	-.01
List order satisfaction it.5	.775	-.01	.923	-.01	.378	-.00
List order satisfaction average	.798	-.01	.693	-.01	.250	.01

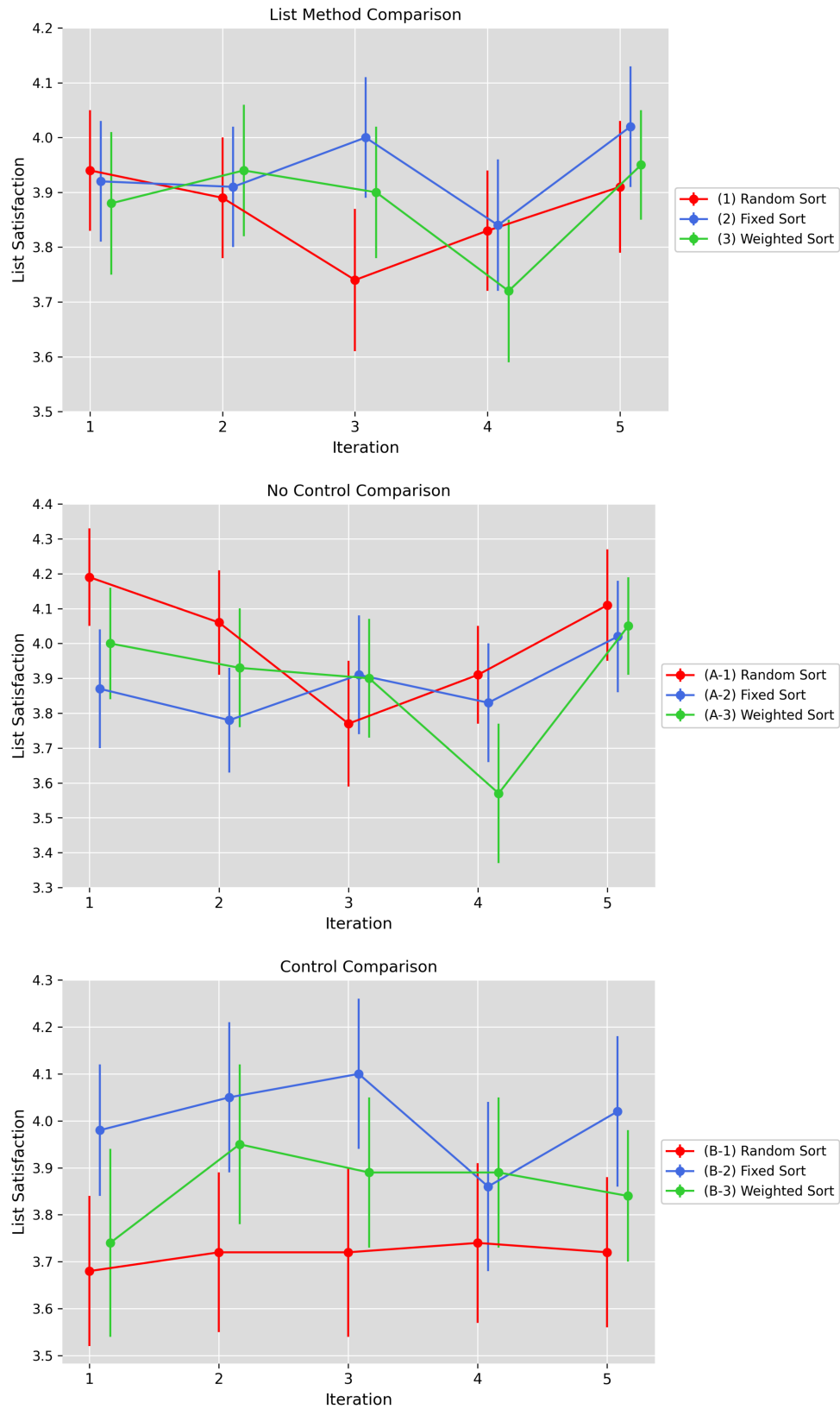


Figure 4.1: Line plot showing how the mean satisfaction score changed per iteration between conditions. Error bars measured in standard error.

## 4.4 Demographic Effects on Evaluation (RQ4)

This section details the results from the analysis performed to gauge if the selected demographic properties affect participant evaluation of the system (RQ4). The first step taken before looking at demographic preferences is to ensure that they are not disproportional between conditions. This was achieved by including these in previous comparison analysis, as they are stored as continuous variables. Based on previous comparative analysis performed between conditions on these metrics in Tables 4.1, B.1, B.2, B.3 and B.5, demographic values are assumed to be normally distributed between conditions. The following test is performed on the whole set of data available, meaning all 300 initial samples.

Table 4.4: Original distribution of samples across the demographic metrics(left) and the re-organized categories used for more even sample sizes..

	Original Data Distribution								Utilized distribution		
<i>Movies Watched</i>	0	1	2	3	4	5	6	7	0to1	2to3	4+
All	8	51	81	49	45	30	14	22	59	130	111
Attention	8	50	75	41	29	27	10	22	58	116	88
<i>Age</i>	18-24	25-34	35-44	45-54	55+				18-34	35-44	45+
All	8	140	85	40	27				148	85	67
Attention	8	119	74	37	24				127	74	61
<i>Gender</i>	Other	Female	Male						Female	Male	
All	3	137	160						137	160	
Attention	3	121	138						121	138	
<i>Streaming Usage</i>	Daily	Weekly+	Quarterly	Hardly					Daily	Rarer	
All	123	162	5	10					123	177	
Attention	109	139	4	10					109	153	

The next step is to investigate the sample size of the demographic metrics to ensure the reliability of the results. Since some options had too low samples to be considered reliably utilized as a categorical variable, the groups were reorganized to provide more even groups with higher samples. The focus was on preserving as much granularity as possible. Original and utilized distribution can be seen in Table 4.4. Alternative errorbar plots can be seen in Appendix B.3.1, Figure B.11 Since streaming usage and gender were reduced to two options, it was decided to use a T-test for pairwise comparison. For domain experience and age with three options, one-way ANOVA with Tukey-HSD post hoc test was selected.

### 4.4.1 Similar System Experience

Table 4.5: Pairwise comparison with t-test on participants who used similar services daily versus rarer. Means are displayed with standard error and are underlined in metrics where objective better values are present. N denotes sample size. Note: † = significant metric, \* $p < .05$ , \*\* $p < .01$ , \*\*\* $p < .001$ .

	$N_{Daily} = 109$	$N_{Rarer} = 139$		
<i>System Usability</i>	Daily Mean <sub>Daily</sub>	Mean <sub>Rarer</sub>	$p$	Cohen's $d$
SUS Score†	<u>74.6±1.89</u>	66.8±1.60	<b>.002**</b>	.40
Wish to use	<u>3.71±0.12</u>	3.55±0.09	.279	.14
Complexity†	<u>2.03±0.12</u>	2.53±0.11	<b>.002**</b>	-.40
Easy to use	<u>4.18±0.10</u>	3.97±0.08	.097	.21
Assistance need	<u>1.74±0.12</u>	1.99±0.11	.145	-.19
Functions integration	<u>3.70±0.10</u>	3.60±0.08	.446	.1
Inconsistencies†	<u>2.24±0.12</u>	2.75±0.11	<b>.003**</b>	-.39
Easy to learn†	<u>4.15±0.10</u>	3.90±0.08	<b>.046*</b>	.26
Cumbersome†	<u>2.21±0.13</u>	2.55±0.11	<b>.048*</b>	-.25
Confidence using	<u>4.07±0.10</u>	3.91±0.09	.198	.17
A lot to learn†	<u>1.75±0.11</u>	2.37±0.11	<b>&gt;.001***</b>	-.49
<i>Participant Activity</i>				
Click list ordering	2.25±0.42	2.06±0.36	.737	.04
Drag&Drop list ordering†	1.55±0.38	0.64±0.21	<b>.028*</b>	.28
Times click was used	2.25±0.42	2.06±0.36	.737	.04
Times drag&drop was used	0.47±0.12	0.27±0.08	.161	.18
Movie index in list	1.62±0.08	1.75±0.06	.182	-.17
Selected list index	1.69±0.19	1.93±0.15	.298	-.13
<i>Recommendation Quality</i>				
List order satisfaction	3.92±0.07	<u>3.93±0.07</u>	.997	0.0
Diversity	<u>3.92±0.09</u>	3.73±0.08	.118	.20
Novelty	<u>3.92±0.11</u>	3.71±0.09	.14	.19
Serendipity	<u>3.71±0.10</u>	3.58±0.09	.329	.13
Relevance	<u>3.39±0.11</u>	3.35±0.10	.745	.04
<i>System Satisfaction</i>				
Recommendation sim†	2.17±0.12	<u>2.65±0.11</u>	<b>.004**</b>	-.37
Difficult selecting movie†	<u>2.46±0.13</u>	2.89±0.11	<b>.009**</b>	-.34
Recommendation Improved	<u>3.41±0.11</u>	3.23±0.09	.211	.16
Felt in control	<u>3.48±0.11</u>	3.23±0.10	.098	.21
Sorting function useful	3.59±0.11	3.45±0.09	.335	.12
<i>Demographic</i>				
Age category†	2.61±0.09	2.93±0.09	<b>.015*</b>	-.31
Gender	1.50±0.05	1.58±0.04	.225	-.16
Movies per week†	4.03±0.19	2.41±0.11	<b>&gt;.001***</b>	.98

Based on the results from the t-test in Table 4.5, participants who used streaming services daily gave a higher SUS score than users who used these services less often. As noted when

analyzing interaction preference in Table 4.2, they were also more likely to utilize drag & drop. No significant variances were found on recommendation quality. While finding the recommendations less similar, they had fewer difficulties finding movies with the system. In correlation with other demographic variables, participants with daily usage of streaming services tended to be younger and watch significantly more movies. To summarize, participants with more system experience evaluated the systems more positively.

#### **4.4.2 Gender**

Full results from the t-test statistical analysis are available in Appendix B.4.1. Very few variances were found when analyzing if gender affected participant evaluation of the system. With the analysis seen in Table B.6, there are only a few metrics with statistically significant variances, which has somewhat high p-values. As such, gender does not appear to have any particular effect on participant evaluation of the system.

### 4.4.3 Age

The one-way ANOVA analysis of age and movie based on the statistical significant values found in the ANOVA Table B.7 in Appendix B.4.1.

Following the one-way ANOVA results, a Tukey-HSD post hoc test was performed with results shown in Table 4.6. The statistically significant results was that recommendation relevance was higher among the 18-34 group than 45+. The youngest segment also felt more in control as opposed to the oldest segment. Both of the younger segments used similar services more often than the oldest. Overall, age seems to have some minor influence on system evaluation, but not particularly strong effects are noted.

Table 4.6: Post hoc Tukey-HSD analysis results from the statistical significant metrics found in the ANOVA Table B.7 regarding age category of participants. Means are displayed with standard error and are underlined in metrics where objective better values are present. Note: \* $p < .05$ , \*\* $p < .01$ , \*\*\* $p < .001$ .

	Group <sub>1</sub>	Group <sub>2</sub>	Mean <sub>1</sub>	Mean <sub>2</sub>	$p$	Cohen's $d$
Assistance need	18-34	35-44	2.09±0.13	<u>1.69±0.13</u>	.078	-.31
	18-34	45+	2.09±0.13	<u>1.66±0.14</u>	.072	-.35
	35-44	45+	1.69±0.13	<u>1.66±0.14</u>	.900	-.03
Relevance	18-34	35-44	<u>3.51±0.10</u>	3.34±0.14	.559	-.15
	18-34	45+	<u>3.51±0.10</u>	2.97±0.14	<b>.008**</b>	-.48
	35-44	45+	<u>3.34±0.14</u>	2.97±0.14	.158	-.32
Recommendation sim	18-34	35-44	<u>2.65±0.12</u>	2.16±0.15	<b>.027*</b>	-.37
	18-34	45+	<u>2.65±0.12</u>	2.46±0.16	.593	-.15
	35-44	45+	2.16±0.15	<u>2.46±0.16</u>	.385	.24
Felt in control	18-34	35-44	<u>3.54±0.10</u>	3.19±0.15	0.106	-.29
	18-34	45+	<u>3.54±0.10</u>	3.07±0.14	<b>.027*</b>	-.43
	35-44	45+	<u>3.19±0.15</u>	3.07±0.14	.793	-.10
Streaming service usage	18-34	35-44	1.63±0.06	1.58±0.08	.866	-.07
	18-34	45+	1.63±0.06	1.89±0.09	<b>.046*</b>	.37
	35-44	45+	1.58±0.08	1.89±0.09	<b>.029*</b>	.45



#### 4.4.4 Analyzing Movie Usage

Based on the statistical significant values found in the ANOVA Table B.7, a Tukey-HSD post hoc test was performed with results shown in Table 4.7. What is noted is that participants who watch fewer movies per week (0-1) had a lower list order satisfaction than the participants who watch more. They also found recommendations to improve less over time, be less surprised with the recommendations, or find them as relevant as participants who watched many movies( 4+) a week. (4+) did also find the sorting function more useful. Overall, domain experienced has a large impact on participant system satisfaction and recommendation quality.

Table 4.7: Post hoc Tukey-HSD analysis results from the statistical significant metrics found in the ANOVA Table B.7 on how many movies participants watched per week. Means are displayed with standard error and are underlined in metrics where objective better values are present. Note: \* $p < .05$ , \*\* $p < .01$ , \*\*\* $p < .001$ .

	Group <sub>1</sub>	Group <sub>2</sub>	Mean <sub>1</sub>	Mean <sub>2</sub>	$p$	Cohen's $d$
Wish to use	0-1	2-3	3.14±0.17	<u>3.58±0.11</u>	.046	.36
	0-1	4+	3.14±0.17	<u>3.85±0.11</u>	.001	.61
	2-3	4+	3.58±0.11	<u>3.85±0.11</u>	.207	.25
Assistance need	0-1	2-3	<u>1.55±0.14</u>	1.78±0.11	.491	.21
	0-1	4+	<u>1.55±0.14</u>	2.22±0.16	<b>.006**</b>	.52
	2-3	4+	<u>1.78±0.11</u>	2.22±0.16	<b>.044*</b>	.32
List order satisfaction	0-1	2-3	3.59±0.10	<u>3.94±0.07</u>	<b>.015*</b>	.44
	0-1	4+	3.59±0.10	<u>4.03±0.07</u>	<b>.002**</b>	.59
	2-3	4+	3.94±0.07	<u>4.03±0.07</u>	.641	.13
Serendipity	0-1	2-3	3.33±0.15	<u>3.67±0.10</u>	.111	.31
	0-1	4+	3.33±0.15	<u>3.76±0.10</u>	<b>.037*</b>	.42
	2-3	4+	3.67±0.10	<u>3.76±0.10</u>	.764	.10
Relevance	0-1	2-3	3.07±0.15	<u>3.29±0.12</u>	.460	.19
	0-1	4+	3.07±0.15	<u>3.57±0.12</u>	<b>.032*</b>	.46
	2-3	4+	3.29±0.12	<u>3.57±0.12</u>	.221	.23
Recommendation Improved	0-1	2-3	3.02±0.14	<u>3.28±0.11</u>	.319	.23
	0-1	4+	3.02±0.14	<u>3.57±0.11</u>	<b>.010*</b>	.53
	2-3	4+	3.28±0.11	<u>3.57±0.11</u>	.152	.26
Felt in control	0-1	2-3	3.07±0.13	<u>3.21±0.12</u>	.720	.12
	0-1	4+	3.07±0.13	<u>3.66±0.12</u>	<b>.008**</b>	.55
	2-3	4+	3.21±0.12	<u>3.66±0.12</u>	<b>.017*</b>	.38
Sorting function useful	0-1	2-3	3.22±0.15	<u>3.47±0.10</u>	.350	.22
	0-1	4+	3.22±0.15	<u>3.72±0.12</u>	<b>.027*</b>	.43
	2-3	4+	3.47±0.10	<u>3.72±0.12</u>	.283	.22
Streaming service usage	0-1	2-3	2.22±0.12	1.64±0.04	<b>.001**</b>	-.83
	0-1	4+	2.22±0.12	1.36±0.06	<b>.001**</b>	-1.16
	2-3	4+	1.64±0.04	1.36±0.06	<b>.001**</b>	-.52



# Chapter 5

## Discussion, Summary & Future Work

### 5.1 Discussion

The principal goal of this thesis was to evaluate if controllability in the multi-lists presentation of recommendations in a similar item scenario is beneficial. This research began by looking at similar studies and found none that have evaluated this approach. However, much was learned from other studies on multi-lists presentation of recommendations and other control implementations, including how demographic metrics may affect evaluation. Findings from other studies were then utilized to develop a prototype for a controllable multi-lists recommender system. This prototype was then evaluated in a large user study, and extensive statistical analysis was performed to address the research question set. The following is a review of the analysis results and findings relating to each research question, with a summary and conclusions at the end.

#### 5.1.1 Evaluating Control (RQ1)

Controllable multi-lists presentation of recommendations was assumed to be beneficial based on other studies on control potentially, [41, 29, 10, 18, 8] and suggestions its utility in the context of multi-list presentation [2, 31]. With controllability in multi-lists presentations having not been evaluated in other studies, many assumptions had to be employed when developing the prototype and rigorously tested to provide a baseline for future research.

Based on the results detailed in Section 4.1, the response to RQ1 is that controllability ap-

appears to have an adverse effect on system usability, with some minor improvements on system satisfaction. As such, it does not appear to improve the user experience for the general participant. It does seem dependent on which list sort method was employed, but no interaction effect was found as evidence to this.

### **5.1.2 Evaluating Different Interaction Methods (RQ2)**

A primary finding when looking at variances in evaluation based on participants' most-used interaction element for control is the lack of engagement by participants. While it was not found as low as in [25] on users selecting recommendation methods, a third of the participants never interacted with the system. Drag& drop was the method chosen the least, with only 23 of 150 participants using this the most, with 76 participants using the click arrows. Most users would prefer to use clickable arrows to reorder the lists if given a choice.

There are some indicators that drag & drop users have a better system evaluation than click users. This may be correlated to the noted finding that participants with higher weekly movie consumption and streaming service usage also had higher enjoyment of the study. However, possibly due to the low sample size, none of these indicators was statistically significant, and no conclusion can therefore be made on the subject. It may be speculated that unfamiliar users trended towards the interactions method they had previous experience with, which may have been the prototype's inferior choice.

Overall, the response to RQ2 is that interaction chosen by a user may affect participants' evaluation of the system, but due to the low sample size, this thesis can make no conclusions on this topic.

### **5.1.3 The Effect of List-sort Methods (RQ3)**

Incorporating control in a recommender system entails some form of system interpretation of user feedback. To this end, three methods were selected. One baseline for investigation if interpretation did not matter, fixed sort, and weighted sort based on literature suggestions[9, 31, 55].

Based on the ANOVA analysis performed on list-order satisfaction, no statistically significant variances between the list sort method were found, neither total nor inter-conditional in

control. No particular variances were found when looking at the other metric categories in the interaction effect analysis of RQ1. The conclusion is that there is no statistical significant discernible differences between the three methods employed in this study and that list order methods may not matter. It may be that the multi-lists presentation of recommendations is enough, that participants do not care about the ordering or notice any differences between the methods, or merely the effect of the study design that influences this.

#### **5.1.4 The Effect of Demographics (RQ4)**

The demographic analysis of the selected demographic properties was performed to detect any of these influenced participants' evaluation of the system. While gender has shown to impact evaluation in other studies [21, 48, 47], both in the movie and other domains, gender had no particular effect on in this context. Age was reported to have mixed effects in other studies [21, 6] and had some minor impact on participants' evaluations in this study.

However, the two metrics that were not found to be evaluated or particularly considered in discovered related literature had the most considerable impact on participation evaluation. System experience and domain experience significantly impacted participants' evaluation of the system, with more experienced users providing a more positive evaluation. As no similar literature was discovered on the topic, it is unknown if this is limited to this study or has broader implications.

#### **5.1.5 Summary of Findings**

Based on the sum of findings, it may be speculated that controllable multi-lists presentation of recommendations is not beneficial to the general user. However, based on the demographic analysis, it may be possible that some niches such as knowledgeable or experienced users would find such a system useful, but more research is needed for this. The following is a summary of the significant findings of this study.

- Controllability in multi lists presentation of recommendations is overall not evaluated to be beneficial to users and has a detrimental effect on system usability. There is, however, evidence that suggests that this might be demographically dependent in terms of

user experience with the domain and similar systems and may therefore be beneficial to a subset consisting of experienced users.

- 13% used drag & drop as their primary interaction method for control, but there are indicators these users had a more positive view of the system. These users also tended to be more familiar with similar systems. The sample size is too small for any indicators to be conclusive.
- No variances on list order satisfaction was found between any of the three used list sort methods. No significant variances were detected when expanding into the general metric set of this study. Based on this, list sort methods evaluated do not appear to significantly impact the user experience as they do not deviate significantly from the baseline.
- Gender and age did not appear to have any particular effect on participant evaluation of the system; however, similar systems and domain experience greatly influenced participants' evaluation of the system.

## **5.2 Limitations**

There are multiple limitations to this thesis. One notable limitation is that the study is run on Mturk and not an existing platform, which may influence the results. Additionally, this thesis only covers the movie domain, so it is uncertain whether this thesis's findings will transfer to other domains such as food recipes or online shopping. The study was also aimed at personal computer users, and a potential discrepancy of results with other platforms such as TV or mobile users was not explored. Finally, the sample size is relatively small in some aspects, particularly when analyzing interaction preferences, which could have resulted in some undiscovered findings.

## 5.3 Future Research

Section 3.2.4 describes how a participant run study tallied 140 individual logged metrics. Not all of these metrics fell within the scope of the research questions of this thesis. An example of unexplored metrics is the index of lists per iterations, which lists participants selected and which movie participants selected to watch later. To enable future research into these topics, the study results are made available for future work. An example of possible analysis is the index of different lists, which are extensively stored in terms of how participants rearranged these lists and how different conditions affected participants' pick of lists and average order. Other metrics include which movie participants tended to choose and preferred to operate in a zoomed-in environment or keep the recommendations in its original size.

Based on results generated through the conducted work, it is made clear that more research is needed for control and list ordering to be beneficial for users in a multi-list environment. More extended studies that measure effect over a longer time, design aspects, and other alternative control techniques are encouraged. The results related to system experience and domain experience as strong influential components to participants' evaluation of the system also warrant more research to verify these findings. A plausible indication from this study is that the general user finds control unfavorable, but the more experienced user might find it appealing. While this study did not find controllability as a positive influence in multi-lists presentation, other alternative implementations may prove it to be beneficial.

## 5.4 Open Science

One of this study's contributions is to make the prototype and study results open to the scientific community as a part of open science. Both of these are available in a Github repository<sup>1</sup>. The shared repository includes the source code for the prototype. Since this data was externally received from the preliminary research authors, Trattner & Jannach [54], explicit consent is needed to receive a copy. It is, however, possible to use the MovieLens database<sup>2</sup> and modify the createDatabase.py script in conjunction with custom similarity calculations located in the repository to recreate the database.

---

<sup>1</sup>Repository location: <https://github.com/Daedalusish/Thesis-Project> last updated 30.11.20

<sup>2</sup>MovieLens dataset location <https://grouplens.org/datasets/movielens/latest/>

The results are divided into multiple categories. At the root level, a folder named "Study result"s contains two versions. The result used in this thesis is available in the "Cleaned data" folder as CSV files organized by condition. The "Raw data" is unprocessed results from the study with more metrics than those available in the .csv files, but due to the nature of Mturk, many entries are incomplete. The `json_to_json.py` script can easily be modified to create new cleaned CSV files with additional metrics if desired.



# Bibliography

- [1] M. Y. H. Al-Shamri. User profiling approaches for demographic recommender systems. *Knowledge-Based Systems*, 100:175–187, 5 2016. ISSN 09507051. doi: 10.1016/j.knosys.2016.03.006.
- [2] O. Alvarado, V. V. Abeele, D. Geerts, and K. Verbert. “I Really Don’t Know What ‘Thumbs Up’ Means”: Algorithmic Experience in Movie Recommender Algorithms. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11748 LNCS, pages 521–541. Springer Verlag, 9 2019. ISBN 9783030293864. doi: 10.1007/978-3-030-29387-1\_{\\_}30. URL [https://doi.org/10.1007/978-3-030-29387-1\\_30](https://doi.org/10.1007/978-3-030-29387-1_30).
- [3] C. Alvino and J. Basilico. Learning a Personalized Homepage, 2015. URL <http://techblog.netflix.com/search/label/recommendations?updated-max=2016-02-12T09:57:00-08:00&max-results=20&start=2&by-date=false>.
- [4] F. Amat, A. Chandrashekar, T. Jebara, and J. Basilico. Artwork personalization at netflix, 2018. URL <https://medium.com/netflix-techblog/artwork-personalization-c589f074ad76>.
- [5] X. Amatriain and J. Basilico. Netflix Recommendations: Beyond the 5 stars (Part 2). *The Netflix Tech Blog*, 2012. URL <http://techblog.netflix.com/2012/06/netflix-recommendations-beyond-5-stars.ht>.
- [6] J. Beel, S. Langer, A. Nürnberger, and M. Genzmehr. The impact of demographics (age and gender) and other user-characteristics on evaluating recommender systems. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 8092 LNCS, pages 396–400, 2013. ISBN 9783642405006. doi: 10.1007/978-3-642-40501-3\_{\\_}45.

- [7] D. Bollen, B. P. Knijnenburg, M. C. Willemsen, and M. Graus. Understanding choice overload in recommender systems. In *RecSys'10 - Proceedings of the 4th ACM Conference on Recommender Systems*, pages 63–70, 2010. ISBN 9781450304429. doi: 10.1145/1864708.1864724. URL <http://www.bbc.co.uk/iplayer>.
- [8] S. Bostandjiev, J. O'Donovan, and T. Höllerer. Tasteweights: A visual interactive hybrid recommender system. In *RecSys'12 - Proceedings of the 6th ACM Conference on Recommender Systems*, pages 35–42, 2012. ISBN 9781450312707. doi: 10.1145/2365952.2365964. URL <http://bit.ly/TasteWeights>.
- [9] C. E. Briguez, M. C. Budán, C. A. Deagustini, A. G. Maguitman, M. Capobianco, and G. R. Simari. Argument-based mixed recommenders and their application to movie suggestion. *Expert Systems with Applications*, 41(14):6467–6482, 10 2014. ISSN 09574174. doi: 10.1016/j.eswa.2014.03.046.
- [10] F. Cena, E. Chiabrando, A. Crevola, M. Deplano, C. Gena, and M. Perrero. A movie timeline for a movie recommender. In *CEUR Workshop Proceedings*, volume 1125, 2013. URL <https://www.themoviedb.org/>.
- [11] X. Chen, P. Zhao, J. Xu, Z. Li, L. Zhao, Y. Liu, V. S. Sheng, and Z. Cui. Exploiting Visual Contents in Posters and Still Frames for Movie Recommendation. *IEEE Access*, 6:68874–68881, 2018. ISSN 21693536. doi: 10.1109/ACCESS.2018.2879971.
- [12] A. Collins, J. Beel, and D. Tkaczyk. One-at-a-time: A Meta-Learning Recommender-System for Recommendation-Algorithm Selection on Micro Level. *Cornell Univesity*, 2018. URL <http://arxiv.org/abs/1805.12118>.
- [13] L. Colucci, P. Doshi, K. L. Lee, J. Liang, Y. Lin, I. Vashishtha, J. Zhang, and A. Jude. Evaluating item-item similarity algorithms for movies. In *Conference on Human Factors in Computing Systems - Proceedings*, volume 07-12-May-, pages 2141–2147, 2016. ISBN 9781450340823. doi: 10.1145/2851581.2892362.
- [14] D. Cosley, S. K. Lam, I. Albert, J. A. Konstan, and J. Riedl. Is seeing believing? How recommender interfaces affect users' opinions. In *Conference on Human Factors in Computing Systems - Proceedings*, pages 585–592, 2003.

- [15] E. D. de Leeuw. *International Handbook of Survey Methodology*. Routledge, 1st edition, 2012. ISBN 9780203843123. doi: 10.4324/9780203843123.
- [16] C. Di Sciascio, V. Sabol, and E. E. Veas. Rank as you go: User-driven exploration of search results. In *International Conference on Intelligent User Interfaces, Proceedings IUI*, volume 07-10-Marc, pages 118–129, 2016. ISBN 9781450341370. doi: 10.1145/2856767.2856797. URL <http://dx.doi.org/10.1145/2856767.2856797>.
- [17] R. Durstenfeld. Algorithm 235: Random permutation. *Communications of the ACM*, 1964. ISSN 15577317. doi: 10.1145/364520.364540.
- [18] M. D. Ekstrand and M. C. Willemsen. Behaviorism is not enough: Better recommendations through listening to users. In *RecSys 2016 - Proceedings of the 10th ACM Conference on Recommender Systems*, pages 221–224, 2016. ISBN 9781450340359. doi: 10.1145/2959100.2959179. URL <http://dx.doi.org/10.1145/2959100.2959179>.
- [19] M. D. Ekstrand, F. M. Harper, M. C. Willemsen, and J. A. Konstan. User perception of differences in recommender algorithms. In *RecSys 2014 - Proceedings of the 8th ACM Conference on Recommender Systems*, pages 161–168, 2014. ISBN 9781450326681. doi: 10.1145/2645710.2645737.
- [20] M. D. Ekstrand, D. Kluver, F. M. Harper, and J. A. Konstan. Letting users choose recommender algorithms: An experimental study. In *RecSys 2015 - Proceedings of the 9th ACM Conference on Recommender Systems*, pages 11–18, 2015. ISBN 9781450336925. doi: 10.1145/2792838.2800195. URL <http://dx.doi.org/10.1145/2792838.2800195>.
- [21] M. D. Ekstrand, M. Tian, I. M. Azpiazu, J. D. Ekstrand, O. Anuyah, D. McNeill, and M. S. Pera. All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. *Conference on Fairness, Accountability and Transparency*, 81:172–186, 2018. doi: 10.18122/B2GM6F. URL <https://dx.doi.org/10.18122/B2GM6F>.
- [22] C. A. Gomez-Uribe and N. Hunt. The netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems*, 6(4):1–19, 2015. ISSN 21586578. doi: 10.1145/2843948.

- [23] A. Gunawardana and G. Shani. Evaluating recommender systems. In *Recommender Systems Handbook, Second Edition*, pages 265–308. Springer, 2015. ISBN 9781489976376. doi: 10.1007/978-1-4899-7637-6{\\_}8.
- [24] F. M. Harper and J. A. Konstan. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, 5(4), 2015. ISSN 21606463. doi: 10.1145/2827872.
- [25] F. M. Harper, F. Xu, H. Kaur, K. Condiff, S. Chang, and L. Terveen. Putting users in control of their recommendations. In *RecSys 2015 - Proceedings of the 9th ACM Conference on Recommender Systems*, pages 3–10, 2015. ISBN 9781450336925. doi: 10.1145/2792838.2800179. URL <http://dx.doi.org/10.1145/2792838.2800179>.
- [26] D. J. Hauser and N. Schwarz. Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, 48(1):400–407, 2016. ISSN 15543528. doi: 10.3758/s13428-015-0578-z.
- [27] C. He, D. Parra, and K. Verbert. Interactive recommender systems: A survey of the state of the art and future research challenges and opportunities. *Expert Systems with Applications*, 56:9–27, 2016. ISSN 09574174. doi: 10.1016/j.eswa.2016.02.013.
- [28] A. R. Hevner, S. T. March, J. Park, and S. Ram. Design science in information systems research. *MIS Quarterly: Management Information Systems*, 28(1):75–105, 2004. ISSN 02767783. doi: 10.2307/25148625.
- [29] R. Hu and P. Pu. Helping users perceive recommendation diversity. In *CEUR Workshop Proceedings*, volume 816, pages 43–50, 2011.
- [30] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich. *Recommender systems: An introduction*, volume 9780521493. Cambridge University Press, Cambridge, illustrate edition, 2010. ISBN 9780511763113. doi: 10.1017/CBO9780511763113.
- [31] M. Jugovac and D. Jannach. Interacting with recommenders-overview and research directions, 2017. ISSN 21606463. URL <https://doi.org/10.1145/3001837>.
- [32] G. Karypis. Evaluation of item-based top-N recommendation algorithms. In *International Conference on Information and Knowledge Management, Proceedings*, pages 247–254, 2001.

- [33] B. P. Knijnenburg, M. C. Willemsen, Z. Gantner, H. Soncu, and C. Newell. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, 22(4-5):441–504, 2012. ISSN 09241868. doi: 10.1007/s11257-011-9118-4.
- [34] V. Krishnan, P. K. Narayanashetty, M. Nathan, R. T. Davies, and J. A. Konstan. Who predicts better? - Results from an online study comparing humans and an online recommender system. In *RecSys'08: Proceedings of the 2008 ACM Conference on Recommender Systems*, pages 211–218, New York, New York, USA, 2008. ACM Press. ISBN 9781605580937. doi: 10.1145/1454008.1454042. URL <http://portal.acm.org/citation.cfm?doid=1454008.1454042>.
- [35] S. Kumar and S. Kumar. An Approach for Recommender System by Combining Collaborative Filtering with User Demographics and Items Genres. *International Journal of Computer Applications*, 128(13):16–24, 2015. doi: 10.5120/ijca2015906724.
- [36] B. Kveton and S. Berkovsky. Minimal interaction content discovery in recommender systems. *ACM Transactions on Interactive Intelligent Systems*, 6(2), 2016. ISSN 21606463. doi: 10.1145/2845090. URL <http://dx.doi.org/10.1145/2845090>.
- [37] J. R. Lewis. The System Usability Scale: Past, Present, and Future. *International Journal of Human-Computer Interaction*, 34(7):577–590, 7 2018. ISSN 15327590. doi: 10.1080/10447318.2018.1455307.
- [38] M. Ludewig and D. Jannach. Evaluation of session-based recommendation algorithms. *User Modeling and User-Adapted Interaction*, 28(4-5):331–390, 2018. ISSN 15731391. doi: 10.1007/s11257-018-9209-6. URL <https://www.dropbox.com/sh/dbzmtq4zhzbj5o9/AACldzQWbw-igKjcPTBI6ZPAa?dl=0>.
- [39] S. M. McNee, S. K. Lam, J. A. Konstan, and J. Riedl. Interfaces for eliciting new user preferences in recommender systems. In *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*, volume 2702, pages 178–187, 2003. doi: 10.1007/3-540-44963-9{\\_}24.
- [40] A. Milton, M. Green, A. Keener, J. Ames, M. D. Ekstrand, and M. S. Pera. StoryTime: Eliciting preferences from children for book recommendations. In *RecSys 2019 - 13th ACM Conference on Recommender Systems*, 2019. ISBN 9781450362436. doi: 10.1145/3298689.3347048.

- [41] T. Nanou, G. Lekakos, and K. Fouskas. The effects of recommendations' presentation on persuasion and satisfaction in a movie recommender system. *Multimedia Systems*, 16(4-5):219–230, 8 2010. ISSN 09424962. doi: 10.1007/s00530-010-0190-0.
- [42] A. Odić, M. Tkalčič, J. F. Tasič, and A. Košir. Relevant context in a movie recommender system: Users' opinion vs. statistical detection. In *CEUR Workshop Proceedings*, volume 889, 2012. URL [www.ldos.si/recommender](http://www.ldos.si/recommender).
- [43] A. Oulasvirta, J. P. Hukkinen, and B. Schwartz. When more is less: The paradox of choice in search engine use. In *Proceedings - 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009*, pages 516–523, 2009. ISBN 9781605584836. doi: 10.1145/1571941.1572030. URL <http://dl.acm.org/citation.cfm?id=1572030>.
- [44] D. Parra, P. Brusilovsky, and C. Trattner. See what you want to see: Visual user-driven approach for hybrid recommendation. In *International Conference on Intelligent User Interfaces, Proceedings IUI*, pages 235–240, 2014. ISBN 9781450321846. doi: 10.1145/2557500.2557542.
- [45] P. Pu, L. Chen, and R. Hu. Evaluating recommender systems from the user's perspective: Survey of the state of the art. *User Modeling and User-Adapted Interaction*, 22(4-5):317–355, 2012. ISSN 09241868. doi: 10.1007/s11257-011-9115-7. URL [www.amazon.com](http://www.amazon.com).
- [46] F. Ricci, L. Rokach, and B. Shapira. Introduction to Recommender Systems Handbook. In *Recommender Systems Handbook*, pages 1–35. Springer, 2011. doi: 10.1007/978-0-387-85820-3{\\_}1.
- [47] M. Rokicki, E. Herder, T. Kuśmierczyk, and C. Trattner. Plate and prejudice: Gender differences in online cooking. In *UMAP 2016 - Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, 2016. ISBN 9781450343701. doi: 10.1145/2930238.2930248.
- [48] A. Said, T. Plumbaum, E. W. De Luca, and S. Albayrak. A Comparison of How Demographic Data Affects Recommendation. In *Adjunct Proceedings of the 2011 Conference on User Modelling, Adaptation and Personalization (UMAP '11)*, 2011. URL [http://www.umap2011.org/proceedings/posters/paper\\_228.pdf](http://www.umap2011.org/proceedings/posters/paper_228.pdf).

- [49] B. Scheibehenne, R. Greifeneder, and P. M. Todd. Can there ever be too many options? A meta-analytic review of choice overload. *Journal of Consumer Research*, 37(3):409–425, 2010. ISSN 00935301. doi: 10.1086/651235.
- [50] M. Sun, C. Li, and H. Zha. Inferring private demographics of new users in recommender systems. In *MSWiM 2017 - Proceedings of the 20th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, 2017. ISBN 9781450351645. doi: 10.1145/3127540.3127566.
- [51] K. Swearingen and R. Sinha. Beyond Algorithms : An HCI Perspective on Recommender Systems. *ACM SIGIR 2001 Workshop on Recommender Systems (2001)*, pages 1–11, 2001. doi: 10.1.1.23.9764. URL [http://www.citeulike.org/user/lrschiff/article/375842%5Cnhttp://citeseer.ist.psu.edu/cache/papers/cs/31330/http:zSzzSzweb.engr.oregonstate.eduzSz%5C~%7B%7DherlockzSzrsw2001zSzfinalzSzfull%5C\\_length%5C\\_paperszSz4%5C\\_swearingenzPz.pdf/swearingen01beyond.pdf](http://www.citeulike.org/user/lrschiff/article/375842%5Cnhttp://citeseer.ist.psu.edu/cache/papers/cs/31330/http:zSzzSzweb.engr.oregonstate.eduzSz%5C~%7B%7DherlockzSzrsw2001zSzfinalzSzfull%5C_length%5C_paperszSz4%5C_swearingenzPz.pdf/swearingen01beyond.pdf).
- [52] K. Swearingen and R. Sinha. Interaction design for recommender systems. In *Designing Interactive Systems*, pages 1–10. In *Designing Interactive Systems 2002*. ACM, 2002. doi: 10.1.1.15.7347. URL <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.15.7347%5Cnhttp://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.15.7347&rep=rep1&type=pdf>.
- [53] N. Tintarev and J. Masthoff. Effective explanations of recommendations: User-centered design. In *RecSys'07: Proceedings of the 2007 ACM Conference on Recommender Systems*, pages 153–156. ACM, 2007. ISBN 9781595937308. doi: 10.1145/1297231.1297259. URL <http://www.csd.abdn.ac.uk/~ntintare/appendix/recSys07.rtf>.
- [54] C. Trattner and D. Jannach. Learning to recommend similar items from human judgments. *User Modeling and User-Adapted Interaction*, 30(1), 2020. ISSN 15731391. doi: 10.1007/s11257-019-09245-4.
- [55] M. Uluyagmur, Z. Cataltepe, and E. Tayfur. Content-based movie recommendation using different feature sets. In *Lecture Notes in Engineering and Computer Science*, volume 1, pages 517–521, 2012. ISBN 9789881925169.

- [56] M. Vozalis and K. Margaritis. Collaborative filtering enhanced by demographic correlation. *AI Symposium on Professional Practice in AI, of the 18th World Computer Congress*, pages 1–10, 2004. doi: 10.1.1.95.8507. URL [http://www.researchgate.net/publication/27382566\\_Collaborative\\_Filtering\\_enhanced\\_by\\_Demographic\\_Correlation/file/d912f50dae2a45ab93.pdf](http://www.researchgate.net/publication/27382566_Collaborative_Filtering_enhanced_by_Demographic_Correlation/file/d912f50dae2a45ab93.pdf).
- [57] Y. Yao and F. Maxwell Harper. Judging similarity: A user-centric study of related item recommendations. In *RecSys 2018 - 12th ACM Conference on Recommender Systems*, volume 100, pages 288–296, Minnesota, USA, 2018. ACM Press. ISBN 9781450359016. doi: 10.1145/3240323.3240351.
- [58] Q. Zhao, G. Adomavicius, F. M. Harper, M. Willemsen, and J. A. Konstan. Toward better interactions in recommender systems: Cycling and serpentine approaches for top-N item lists. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, pages 1444–1453, 2017. ISBN 9781450343350. doi: 10.1145/2998181.2998211. URL <http://dx.doi.org/10.1145/2998181.2998211>.







# **Appendix A**

## **Screenshots**

This appendix contains screenshots from each step of the Mturk study carried out in this thesis.

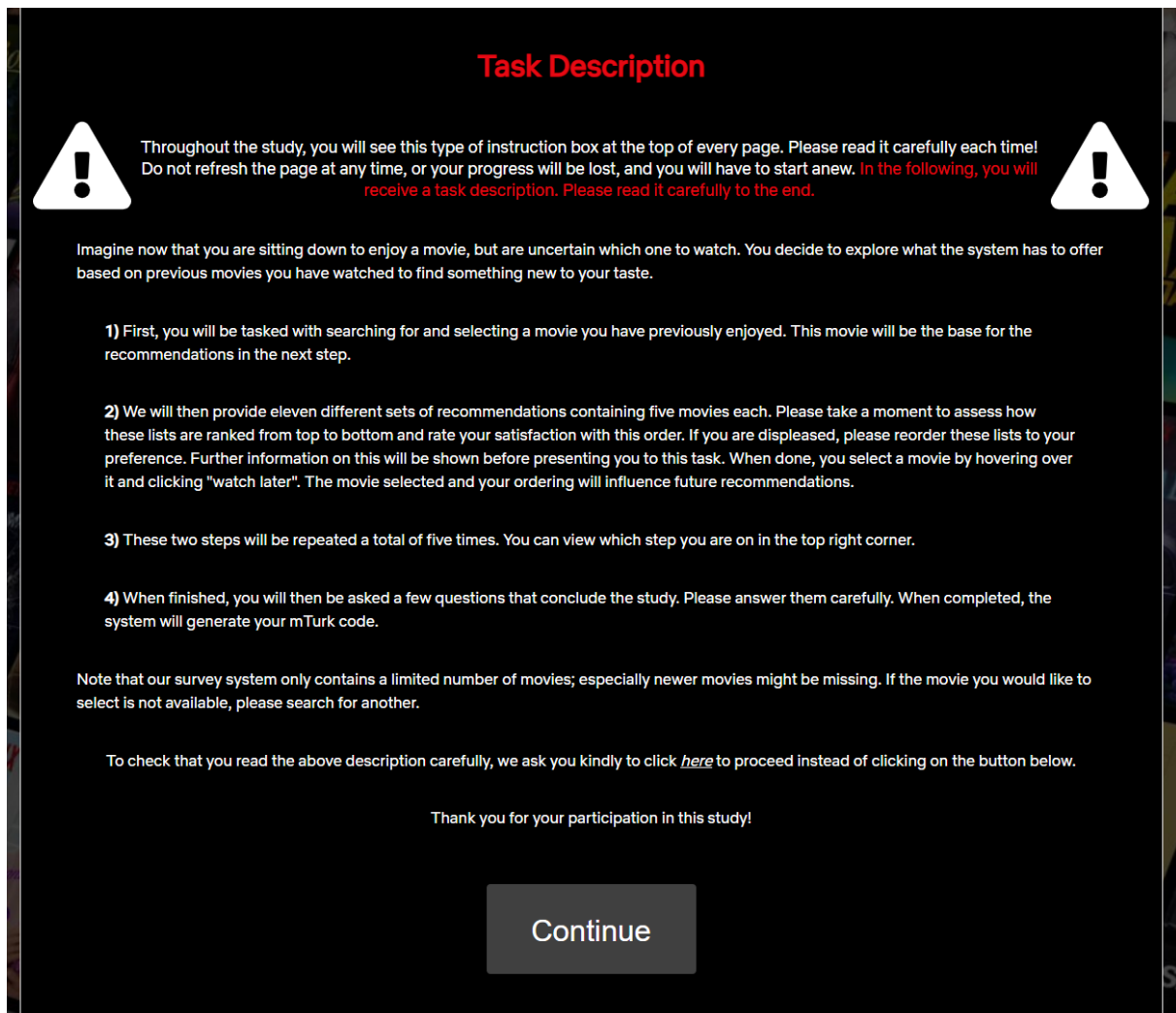


Figure A.1: Screenshot showing the introductions provided to participants as the first step of the study.

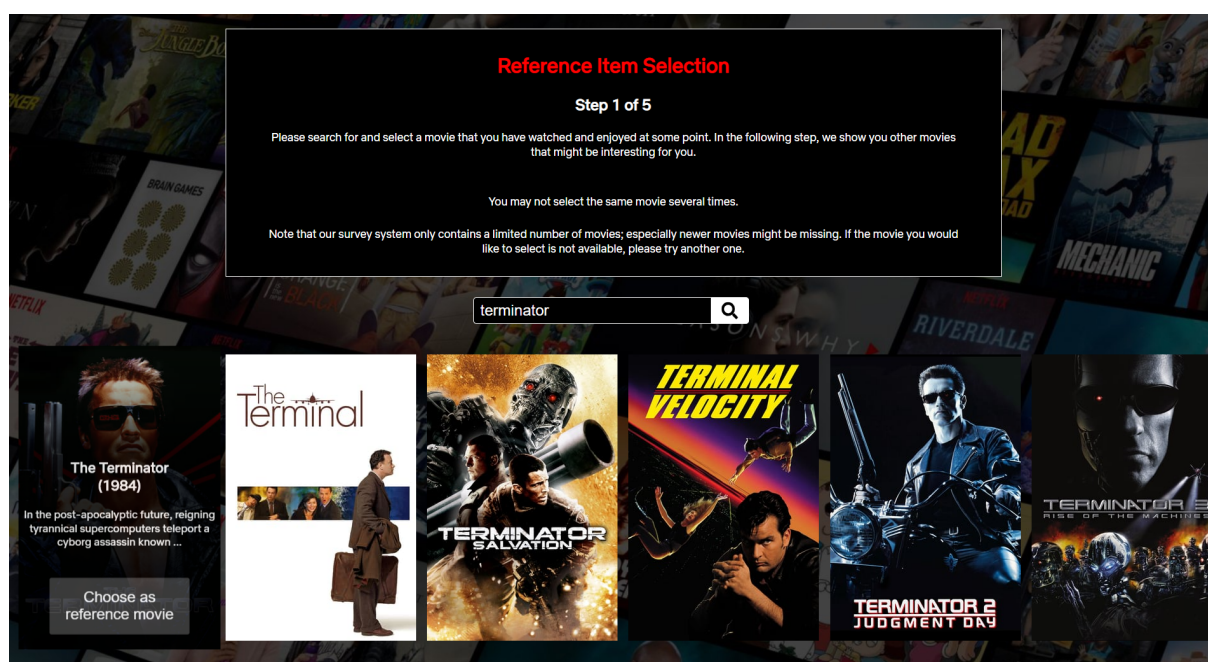



Figure A.2: Screenshot showing the search screen. This the search results display, the evidence provided and how participants select movies


## Instructions

**Please read these instructions carefully before you continue. A summary will also be present at the top of every browsing session.**

Based on the movie you selected, several sets of recommendations will be presented to you. Please inspect the recommendation lists in the next screen and rate how you perceive their ordering from top to bottom before performing any actions. **You cannot move the lists or select a movie before you do this!** After this, you are now tasked with rearranging them to an order that fits your preference. Ordering can be done in two ways. Either by dragging the list or clicking the arrow buttons as illustrated below:



**Method 1 seen above:** By clicking and holding the left-click down while hovering over the "↕" symbol next to a list, you can drag the list to move it down or up the order. The doublesided arrow will be grey when you hover over it and red when you are dragging it.



**Method 2 seen above:** Each list has two sets of arrows - Up and down. By performing a single left-click on an arrow, the list will move into the arrow's direction.

**To summarize:** First, evaluate the list order presented to you and rate their order. Then use either or both ordering methods described above to rearrange the list to more closely fit what you would consider being the best order. When both these tasks are done, hover over a movie and click the "Watch later" button to continue. You will not be able to proceed if you do not rate the selection.

Figure A.3: Supplement introductions added for condition B to clarify and showcase instructions. The images are animated .gifs showcasing how to order.

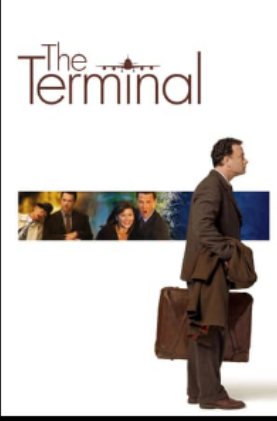
**Ordering Recommendations Provided**
**Step 1 of 5**

Based on the movie you selected as a reference movie in the previous step, which is shown directly below, several sets of recommendations are presented. Please inspect the recommendation lists below and rate how the selection as a whole is ranked from top to bottom. Top ranking indicates that the selection has many movies you find similar. After rating, you may now reorder the lists to your preference. This can be done by dragging the handle icon "↕" or by clicking the arrows to move one position at the time. When done, select a movie that you would like to watch by hovering over one. The movie you select and your ordering of the lists will influence future recommendations.

Your Reference Movie:

## The Terminal

**(2004)**  
 Viktor Navorski is a man without a country; his plane took off just as a coup d'etat exploded in his homeland, leaving it in shambles, and now he's stranded at Kennedy Airport, where he's holding a passport that nobody recognizes. While quarantined in the transit lounge until authorities can figure out what to do with him, Viktor simply goes on living – and courts romance with a beautiful flight attendant.  
 Stars: Tom Hanks, Catherine Zeta-Jones, Stanley Tucci, Chi McBride  
 Director: Steven Spielberg  
 Genres: Comedy, Drama



Please report on a scale from very unsatisfied (1) to very satisfied (5) on how you happy you are with the current list order, from top to bottom.

Very Unsatisfied
1
2
3
4
Very Satisfied

Zoom Out

Here are your movie recommendations, based on...











	Title				
↕					
↕					

Figure A.4: A screenshot of the browse screen where participants are given their recommendations, in this case the control condition (B). Note that recommendation continue below and page is zoomed out to fit content in view



## Final Questionnaire

As a final step, please fill out the questionnaire below. "\*" denotes required fields

**1. To what degree do you agree with the following statements? \***

	Completely Disagree(1)	(2)	(3)	(4)	Completely Agree(5)
I think that I would like to use this system frequently.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I found the system unnecessarily complex.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I thought the system was easy to use.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I think that I would need the support of a technical person to be able to use this system.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I found the various functions in this system well integrated.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I thought there was too much inconsistency in this system	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I would imagine that most people would learn to use this system very quickly.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I found the system very cumbersome to use.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I felt very confident using the system.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I needed to learn a lot of things before I could get going with this system.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Searching for similar movies took a lot of time.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Deciding between a large number of options was difficult.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The recommendations improved over time.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I felt in control of the recommendations.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The sorting functions were useful.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Fill in on the scale the answer to 'one plus two'.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The movies presented to me were diverse.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I was able to discover new movies.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I was pleasantly suprised by the recommendations.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
All of the presented movies were similar to my reference movies.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

**2. Which of the following statements best describes your use of online movie services (e.g, Netflix, IMDB, etc.)? \***

.....

**3. On average, on how many days per week do you watch a movie? \***

.....

**4. Your age: \***

.....

**5. Your gender: \***

.....

Figure A.5: Screenshot of the final survey given to participants, zoomed out to fit more on screen.



# Appendix B

## Statistics

This appendix chapter contains statistics not included in the result chapter. It is organized by which research question the content pertain to, with sub categories on context.

### B.1 RQ1

#### B.1.1 Errorbar Plots RQ1

These are errorbar comparison plots between conditions for RQ1 with the mean value and standard error for each metric.

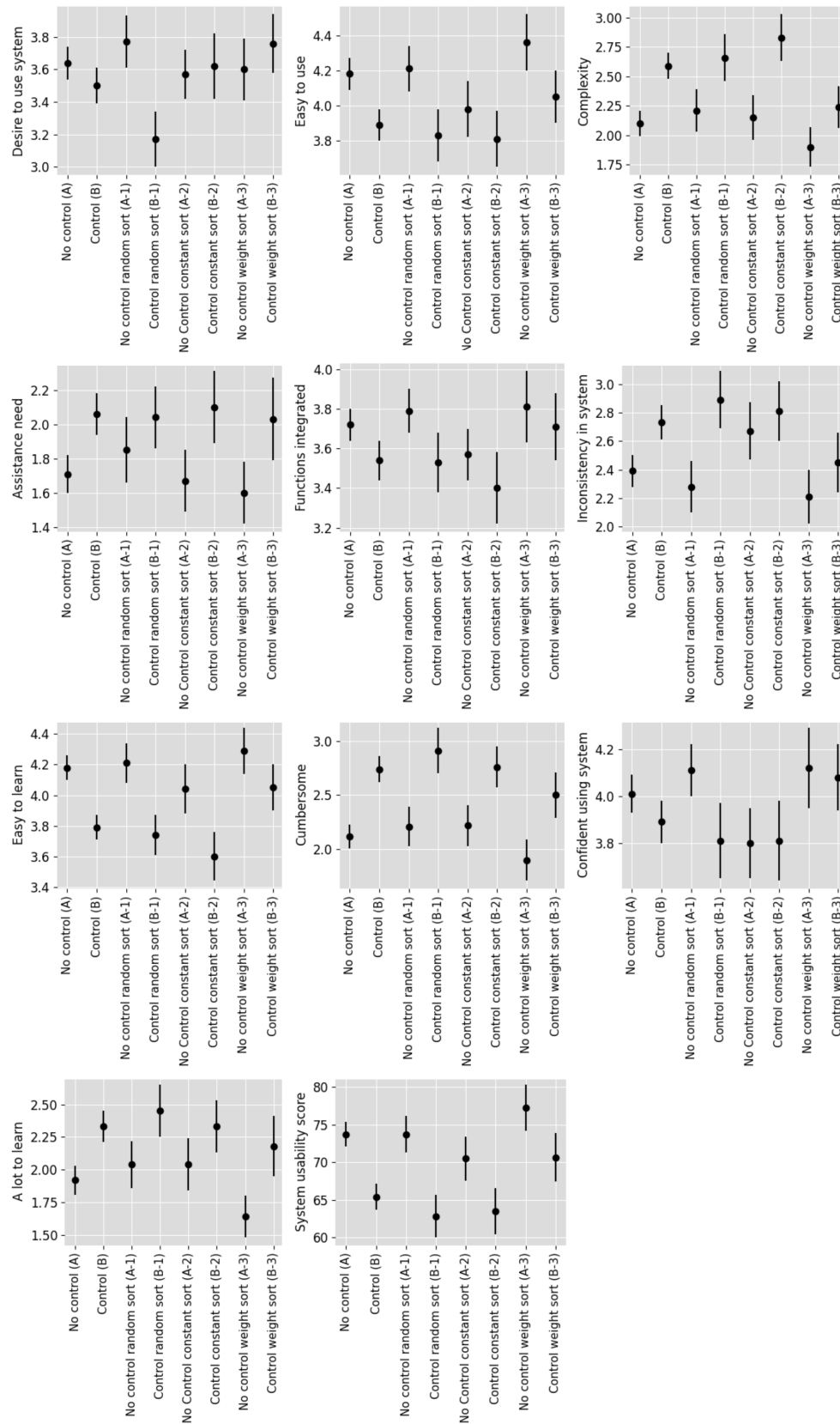


Figure B.1: System usability metrics used in RQ1 in Section 4.1. Comparison plot with mean values and standard error.

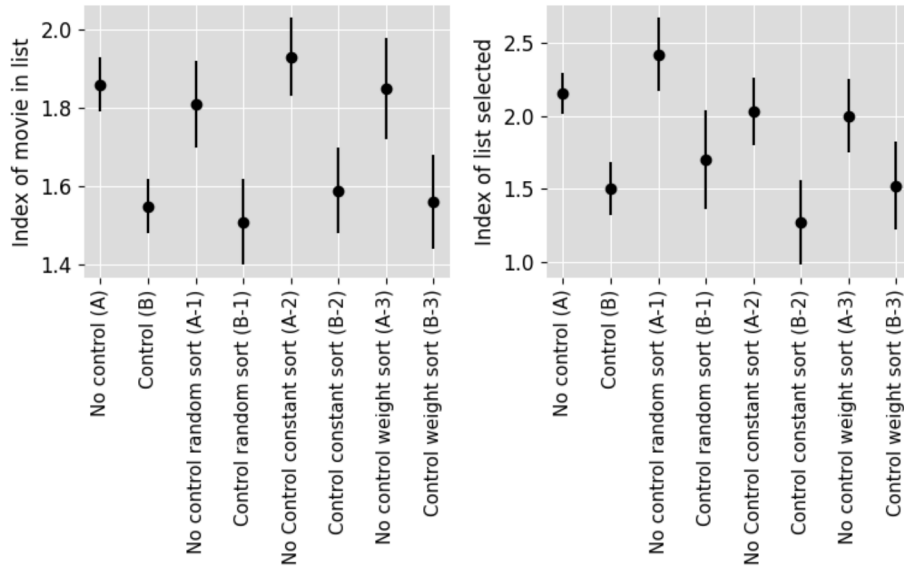


Figure B.2: Activity metrics used in RQ1 in Section 4.1. Comparison plot with mean values and standard error.

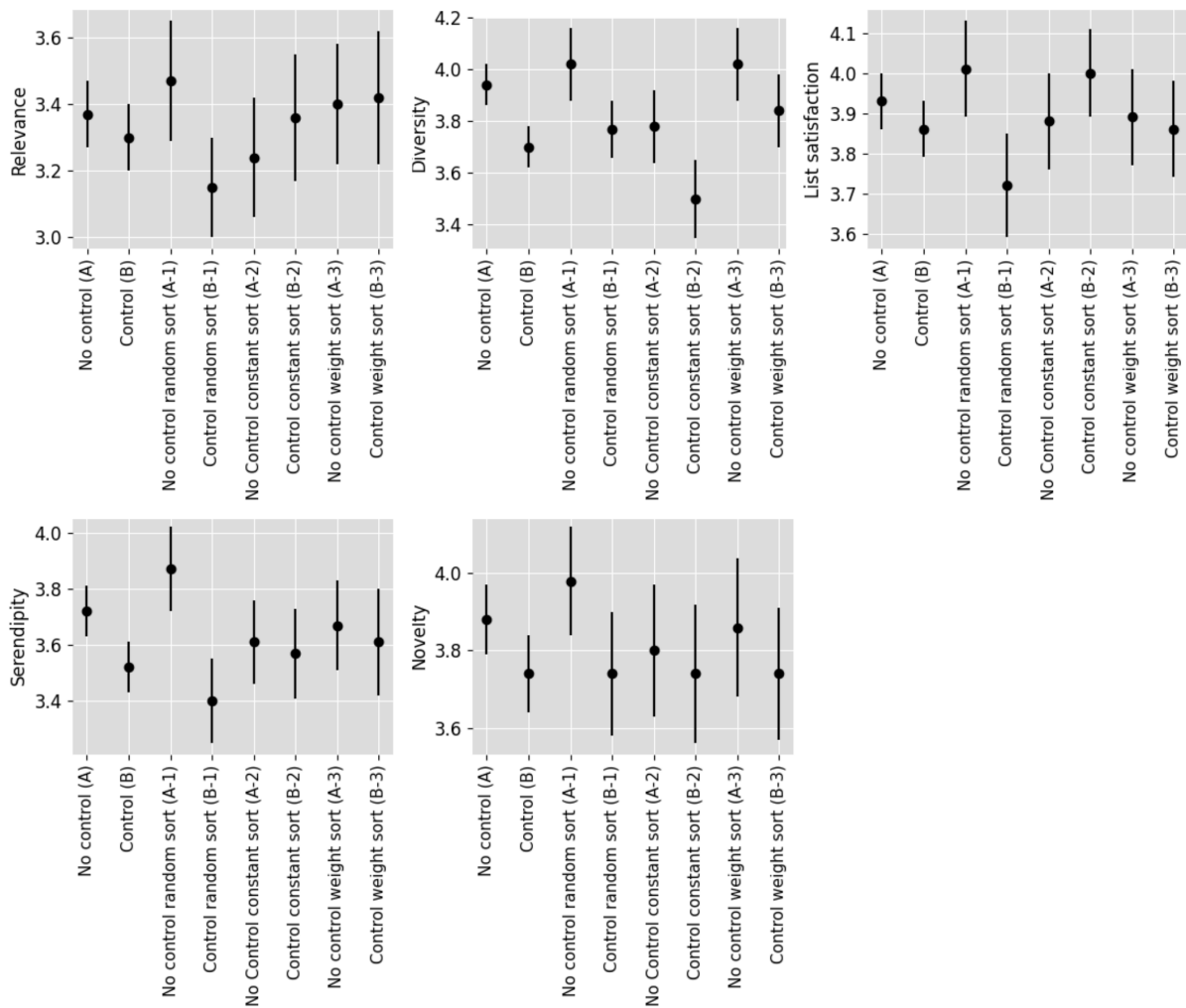


Figure B.3: Recommendation quality metrics used in RQ1 in Section 4.1. Comparison plot with mean values and standard error.

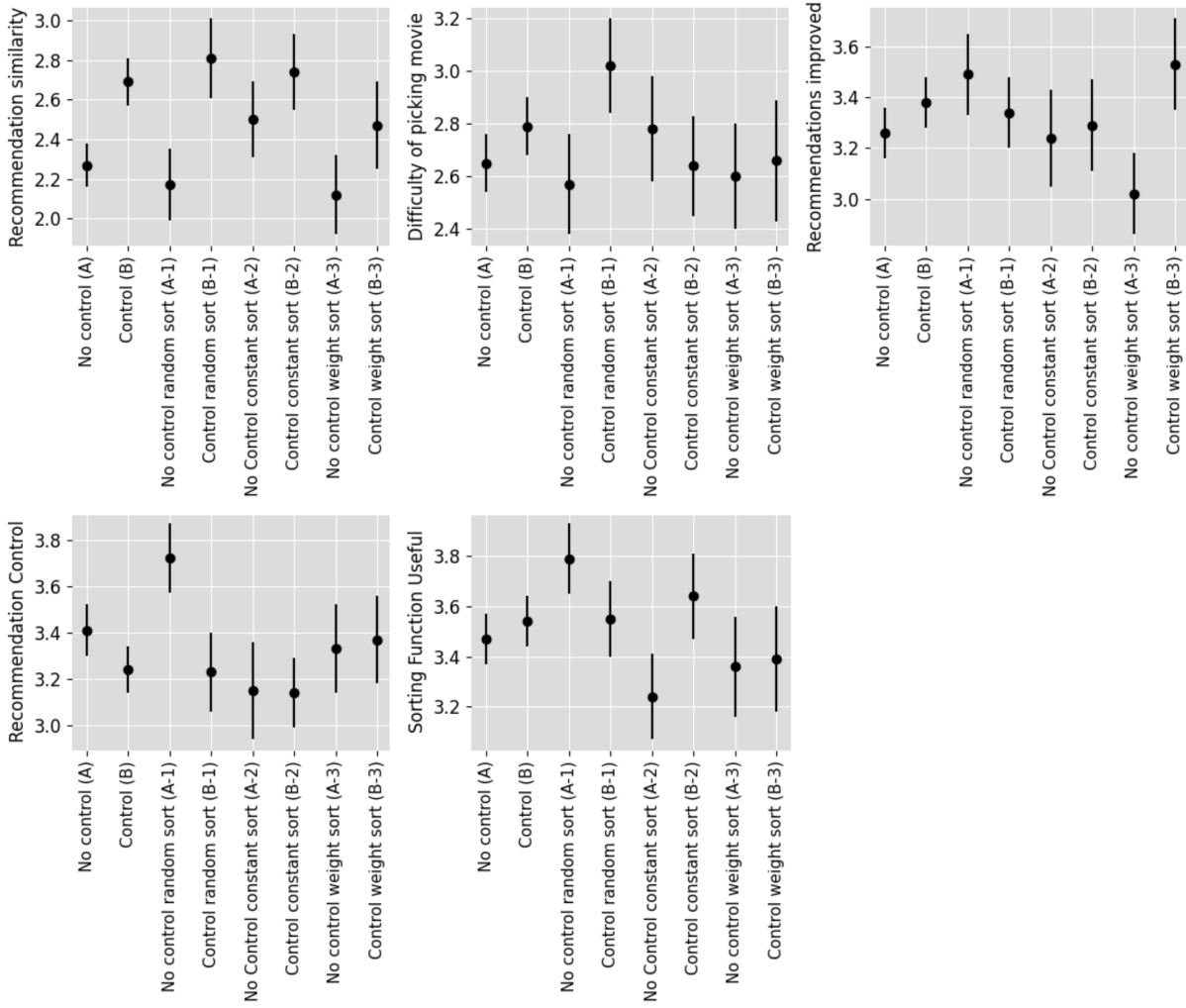


Figure B.4: System satisfaction metrics used in RQ1 in Section 4.1. Comparison plot with mean values and standard error.

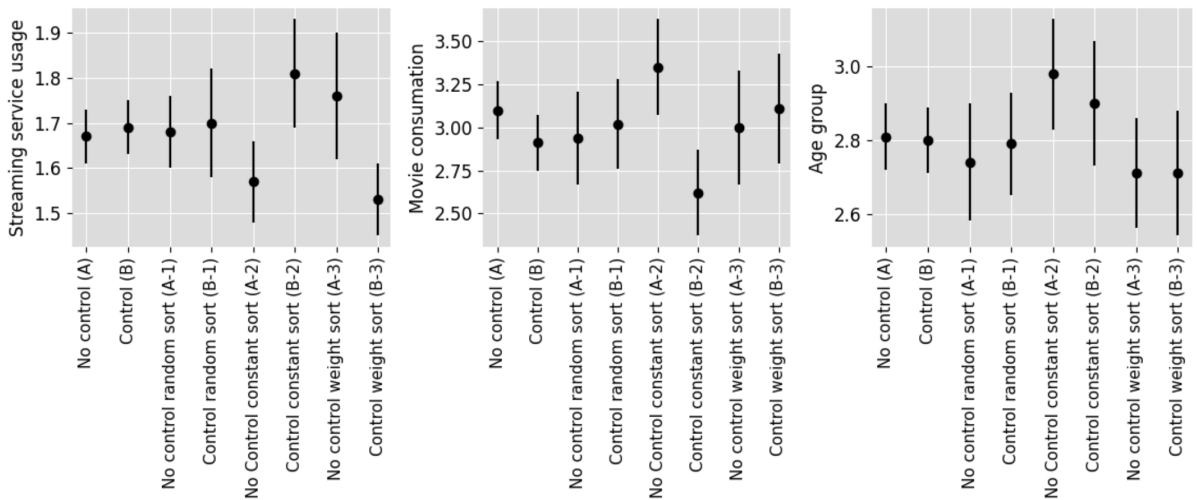


Figure B.5: Demographic metrics in Section 4.1. Comparison plot with mean values and standard error.

### **B.1.2 Sub Condition t-tests for RQ1**

This section contains additional pairwise t-tests performed on sub-conditions for RQ1.

Table B.1: Pairwise comparison with t-test between no control and control when list order is randomized. Means are displayed with standard error and are underlined in metrics where objective better values are present. N denotes sample size. Note: † = significant metric, \* $p < .05$ , \*\* $p < .01$ , \*\*\* $p < .001$ .

	$N_{A-1} = 47$		$N_{B-1} = 47$	
<i>System Usability</i>	Mean <sub>A-1</sub>	Mean <sub>B-1</sub>	<i>p</i>	Cohen's <i>d</i>
SUS Score†	<u>73.72±2.42</u>	62.82±2.85	<b>.004**</b>	.60
Wish to use†	<u>3.77±0.16</u>	3.17±0.17	<b>.013*</b>	.52
Complexity	<u>2.21±0.18</u>	2.66±0.20	.102	-.34
Easy to use	<u>4.21±0.13</u>	3.83±0.15	.050	.41
Assistance need	<u>1.85±0.19</u>	2.04±0.18	.469	-.15
Functions integration	<u>3.79±0.11</u>	3.53±0.15	.175	.28
Inconsistencies†	<u>2.28±0.18</u>	2.89±0.20	<b>.024*</b>	-.47
Easy to learn†	<u>4.21±0.13</u>	3.74±0.13	<b>.011*</b>	.53
Cumbersome†	<u>2.21±0.18</u>	2.91±0.21	<b>.013*</b>	-.52
Confidence using	<u>4.11±0.11</u>	3.81±0.16	.134	.31
A lot to learn	<u>2.04±0.18</u>	2.45±0.20	.135	-.31
<i>Participant Activity</i>				
Movie index in list	1.81±0.11	1.51±0.11	.063	.39
Selected list index	2.42±0.25	1.7±0.34	.096	.35
<i>Recommendation Quality</i>				
List order satisfaction	<u>4.01±0.12</u>	3.72±0.13	.098	.35
Diversity	<u>4.02±0.14</u>	3.77±0.11	.158	.29
Novelty	<u>3.98±0.14</u>	3.74±0.16	.264	.23
Serendipity†	<u>3.87±0.15</u>	3.40±0.15	<b>.031*</b>	.45
Relevance	<u>3.47±0.18</u>	3.15±0.15	.175	.28
<i>System Satisfaction</i>				
Recommendation sim†	2.17±0.18	<u>2.81±0.20</u>	<b>.018*</b>	-.50
Difficult selecting movie	<u>2.57±0.19</u>	3.02±0.18	.092	-.35
Recommendation Improved	<u>3.49±0.16</u>	3.34±0.14	.495	.14
Felt in control†	<u>3.72±0.15</u>	3.23±0.17	<b>.034*</b>	.44
Sorting function useful	<u>3.79±0.14</u>	3.55±0.15	.253	.24
<i>Demographic</i>				
Age category	2.74±0.16	2.79±0.14	.841	-.04
Streaming service usage	1.68±0.08	1.70±0.12	.879	-.03
Gender	1.68±0.07	1.57±0.08	.332	.20
Movies per week	2.94±0.27	3.02±0.26	.823	-.05

Table B.2: Pairwise comparison with t-test between no control and control when list order is fixed. Means are displayed with standard error and are underlined in metrics where objective better values are present. N denotes sample size. Note: † = significant metric, \* $p < .05$ , \*\* $p < .01$ , \*\*\* $p < .001$ .

	$N_{A-2} = 46$		$N_{B-2} = 42$	
<i>System Usability</i>	Mean <sub>A-2</sub>	Mean <sub>B-2</sub>	$p$	Cohen's $d$
SUS Score	<u>7.49±2.91</u>	63.51±3.08	.103	.35
Wish to use	<u>3.57±0.15</u>	<u>3.62±0.20</u>	.830	-.05
Complexity†	<u>2.15±0.19</u>	<u>2.83±0.20</u>	<b>.013*</b>	-.54
Easy to use	<u>3.98±0.16</u>	3.81±0.16	.455	.16
Assistance need	<u>1.67±0.18</u>	2.10±0.21	.128	-.33
Functions integration	<u>3.57±0.13</u>	3.40±0.18	.468	.16
Inconsistencies	<u>2.67±0.20</u>	2.81±0.21	.643	-.10
Easy to learn†	<u>4.04±0.16</u>	3.60±0.16	<b>.049*</b>	.42
Cumbersome†	<u>2.22±0.19</u>	2.76±0.19	<b>.047*</b>	-.43
Confidence using	3.80±0.15	<u>3.81±0.17</u>	.982	-.00
A lot to learn	<u>2.04±0.20</u>	2.33±0.20	.312	-.22
<i>Participant Activity</i>				
Movie index in list†	1.93±0.10	1.59±0.11	<b>.024*</b>	-.49
Selected list index†	2.03±0.23	1.27±0.29	<b>.038*</b>	-.45
<i>Recommendation Quality</i>				
List order satisfaction	3.88±0.12	<u>4.00±0.11</u>	.475	-.15
Diversity	<u>3.78±0.14</u>	3.50±0.15	.174	.29
Novelty	<u>3.80±0.17</u>	3.74±0.18	.789	.06
Serendipity	<u>3.61±0.15</u>	3.57±0.16	.864	.04
Relevance	3.24±0.18	<u>3.36±0.19</u>	.651	-.10
<i>System Satisfaction</i>				
Recommendation sim	2.50±0.19	<u>2.74±0.19</u>	.380	-.19
Difficult selecting movie	<u>2.78±0.20</u>	2.64±0.19	.611	.11
Recommendation Improved	3.24±0.19	<u>3.29±0.18</u>	.861	-.04
Felt in control	<u>3.15±0.21</u>	3.14±0.15	.971	.01
Sorting function useful	3.24±0.17	<u>3.64±0.17</u>	.095	-.36
<i>Demographic</i>				
Age category	2.98±0.15	2.90±0.17	.746	.07
Streaming service usage	1.57±0.09	1.81±0.12	.096	-.36
Gender	1.41±0.07	1.60±0.08	.090	-.37
Movies per week	3.35±0.28	2.62±0.25	.057	.41

Table B.3: Pairwise comparison with t-test between no control and control when list order is weighted. Means are displayed with standard error and are underlined in metrics where objective better values are present. N denotes sample size. Note: † = significant metric, \* $p < .05$ , \*\* $p < .01$ , \*\*\* $p < .001$ .

	$N_{A-3} = 42$		$N_{B-3} = 38$	
<i>System Usability</i>	Mean <sub>A-2</sub>	Mean <sub>B-2</sub>	$p$	Cohen's $d$
SUS Score	<u>77.26±3.04</u>	7.66±3.20	.139	.33
Wish to use	<u>3.60±0.19</u>	<u>3.76±0.18</u>	.528	-.14
Complexity	<u>1.90±0.17</u>	2.24±0.18	.191	-.30
Easy to use	<u>4.36±0.16</u>	4.05±0.15	.160	.32
Assistance need	<u>1.60±0.18</u>	2.03±0.24	.144	-.33
Functions integration	<u>3.81±0.18</u>	3.71±0.17	.692	.09
Inconsistencies	<u>2.21±0.19</u>	2.45±0.21	.420	-.18
Easy to learn	<u>4.29±0.15</u>	4.05±0.15	.263	.25
Cumbersome†	<u>1.90±0.19</u>	2.50±0.21	<b>.036*</b>	-.48
Confidence using	<u>4.12±0.17</u>	4.08±0.14	.861	.04
A lot to learn	<u>1.64±0.16</u>	2.18±0.23	.052	-.44
<i>Participant Activity</i>				
Movie index in list	1.85±0.13	1.56±0.12	.122	.35
Selected list index	2.00±0.25	1.52±0.30	.215	.28
<i>Recommendation Quality</i>				
List order satisfaction	<u>3.89±0.12</u>	3.86±0.12	.872	.04
Diversity	<u>4.02±0.14</u>	3.84±0.14	.366	.20
Novelty	<u>3.86±0.18</u>	3.74±0.17	.627	.11
Serendipity	<u>3.67±0.16</u>	3.61±0.19	.802	.06
Relevance	3.40±0.18	<u>3.42±0.20</u>	.952	-.01
<i>System Satisfaction</i>				
Recommendation sim	2.12±0.20	<u>2.47±0.22</u>	.240	-.26
Difficult selecting movie	2.60±0.20	<u>2.66±0.23</u>	.838	-.05
Recommendation Improved†	3.02±0.16	<u>3.53±0.18</u>	<b>.038*</b>	-.47
Felt in control	3.33±0.19	<u>3.37±0.19</u>	.897	-.03
Sorting function useful	3.36±0.20	<u>3.39±0.21</u>	.896	-.03
<i>Demographic</i>				
Age category	2.71±0.15	2.71±0.17	.987	.00
Streaming service usage	1.76±0.14	1.53±0.08	.151	.32
Gender	1.50±0.09	1.53±0.08	.825	-.05
Movies per week	3.00±0.33	3.11±0.32	.821	-.05



### B.1.3 Interaction Effect Analysis RQ1

This section contains the one-way ANOVA performed between list-sort methods. It is used for evaluating if an interaction effect is present for RQ1 and if list sort methods affected other metrics besides list order satisfaction in RQ3.

Table B.4: Tukey-HSD post hoc test result for statistical significant values found in the one-way ANOVA Table B.5

<i>Post-hoc on list-sort methods in all control conditions (1, 2, 3)</i>						
	Group		Group Mean		<i>p</i>	Cohen's <i>d</i>
	Group <sub>1</sub>	Group <sub>2</sub>	Mean <sub>1</sub>	Mean <sub>2</sub>		
SUS Score	Random	Fixed	<u>68.3±1.94</u>	67.2±2.13	.900	-.06
	Random	Weighted	68.3±1.94	<u>74.1±2.22</u>	.120	.30
	Fixed	Weighted	67.2±2.13	<u>74.1±2.22</u>	.060	.35
<i>Post-hoc on list-sort methods in no control conditions (A-1, A-2, A-3)</i>						
Sorting function useful	Random	Fixed	<u>3.79±0.14</u>	3.24±0.17	.051	-.53
	Random	Weighted	<u>3.79±0.14</u>	3.36±0.20	.170	-.38
	Fixed	Weighted	3.24±0.17	<u>3.36±0.20</u>	.860	.10
Gender	Random	Fixed	<u>1.68±0.07</u>	1.41±0.07	<b>.030*</b>	-.55
	Random	Weighted	<u>1.68±0.07</u>	1.50±0.09	.220	-.35
	Fixed	Weighted	1.41±0.07	<u>1.50±0.09</u>	.680	.17

Table B.5: One-Way ANOVA on list conditions. Means are displayed with standard error and are underlined in metrics where objective better values are present. Note: † = significant metric, \* $p < .05$ , \*\* $p < .01$ , \*\*\* $p < .001$ .

	List Sort (1, 2, 3)		No Control (A-1, A-2, A-3)		Control (B-1, B-2, B-3)	
	$N_1 = 94$		$N_A - 1 = 47$		$N_B - 1 = 47$	
	$N_2 = 88$		$N_A - 2 = 46$		$N_B - 2 = 42$	
	$N_3 = 80$		$N_A - 3 = 52$		$N_B - 3 = 38$	
<i>System Usability</i>	$p$	$\omega^2$	$p$	$\omega^2$	$p$	$\omega^2$
SUS Score	<b>.050*</b>	.02	.240	.01	.150	.01
Wish to use	.500	.00	.660	-.01	.060	.03
Complexity	.070	.01	.460	.00	.100	.02
Easy to use	.120	.01	.190	.01	.470	.00
Assistance need	.760	-.01	.610	-.01	.970	-.02
Functions integration	.210	.00	.400	.00	.450	.00
Inconsistencies	.130	.01	.190	.01	.280	.00
Easy to learn	.070	.01	.480	.00	.090	.02
Cumbersome	.150	.01	.420	.00	.350	.00
Confidence using	.170	.01	.220	.01	.420	.00
A lot to learn	.190	.01	.230	.01	.670	-.01
<i>Participant Activity</i>						
Total list ordering	.270	.00	N/A	N/A	.240	.01
Click list ordering	.710	-.01	N/A	N/A	.750	-.01
Drag&Drop list ordering	.160	.01	N/A	N/A	.170	.01
Times click was used	.710	-.01	N/A	N/A	.750	-.01
Times drag&drop was used	.190	.00	N/A	N/A	.210	.01
Movie index in list	.660	.00	.750	-.01	.890	-.01
Selected list index	.330	.00	.390	.00	.610	-.01
<i>Recommendation Quality</i>						
Diversity	.080	.01	.370	.00	.190	.01
Novelty	.850	-.01	.730	-.01	1.00	-.02
Serendipity	.940	-.01	.430	.00	.640	-.01
Relevance	.780	-.01	.650	-.01	.520	-.01
<i>System Satisfaction</i>						
Recommendation sim	.270	.00	.300	.00	.490	.00
Difficult selecting movie	.680	.00	.710	-.01	.290	.00
Recommendation Improved	.570	.00	.170	.01	.590	-.01
Felt in control	.160	.01	.080	.02	.660	-.01
Sorting function useful†	.180	.01	<b>.049*</b>	.03	.620	-.01
<i>Demographic</i>						
Streaming service usage	.960	-.01	.640	-.01	.440	.00
Movies per week	.920	-.01	.380	.00	.210	.01
Age category	.310	.00	.420	.00	.700	-.01
Gender†	.190	.01	<b>.040*</b>	.03	.840	-.01

## B.2 RQ2

This section contains appendix material relating to RQ2 analysis performed

### B.2.1 Errorbar Plots for RQ2

These are errorbar comparison plots between interaction methods for RQ2 with the mean value and standard error for each metric.

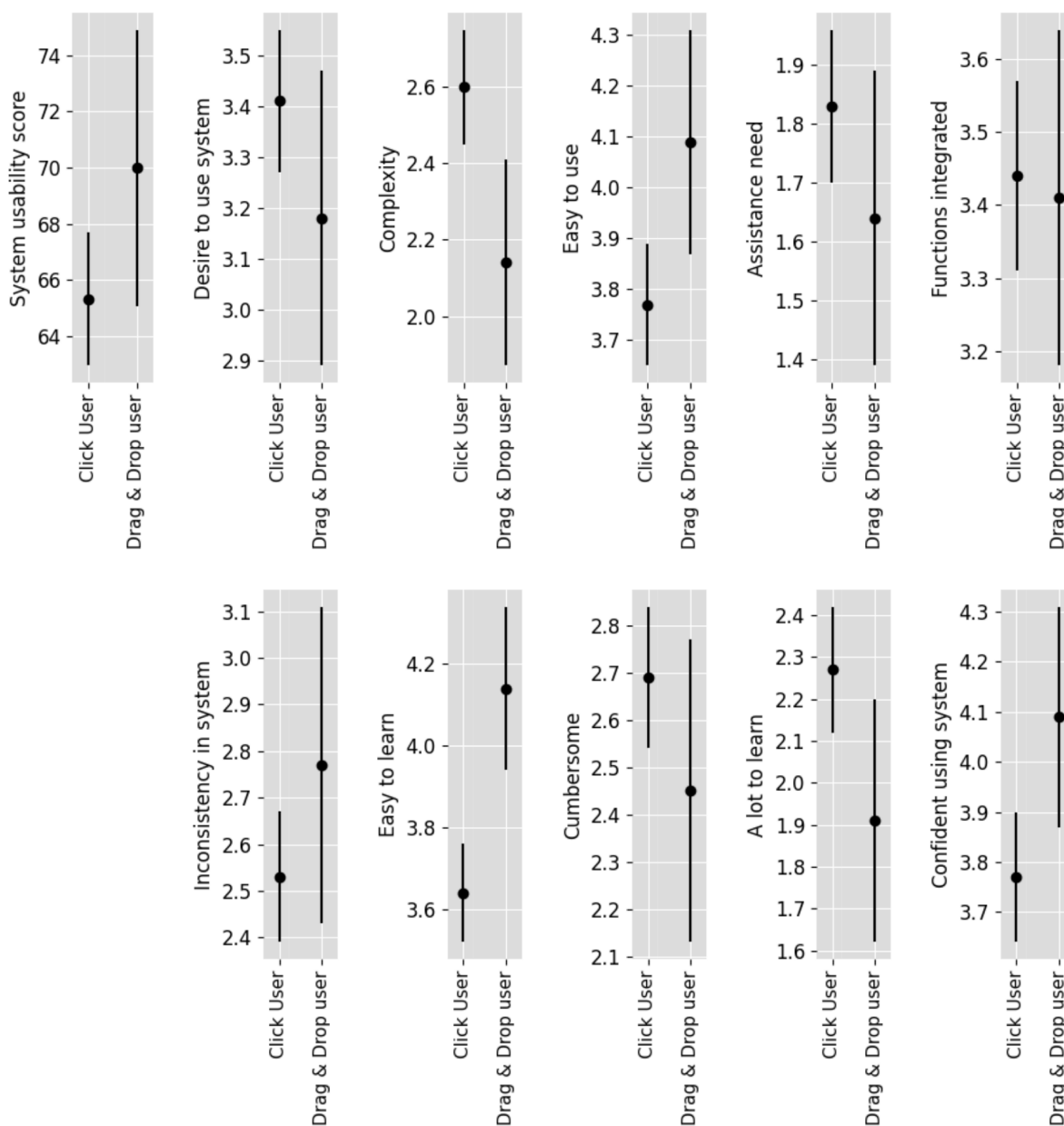


Figure B.6: System metrics used for RQ2 in Section 4.2. Comparison plot with mean values and standard error.

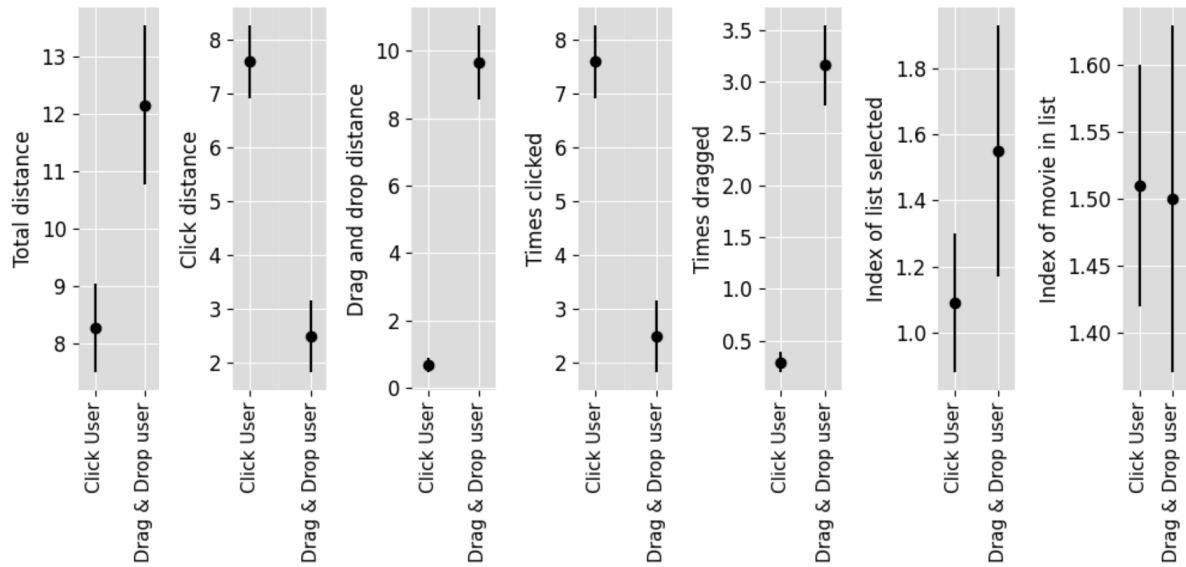


Figure B.7: Activity metrics used for RQ2 in Section 4.2. Comparison plot with mean values and standard error.

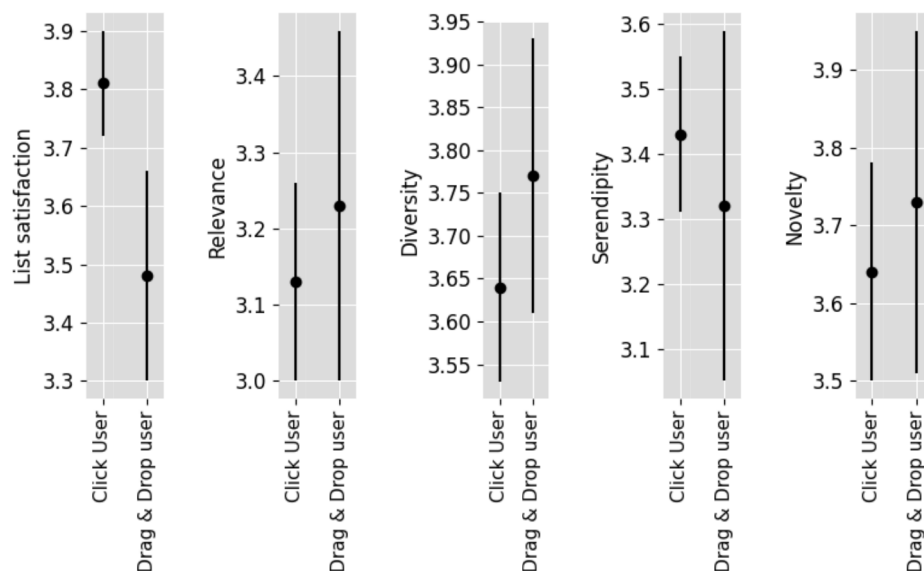


Figure B.8: Recommendation Quality metrics used for RQ2 in Section 4.2. Comparison plot with mean values and standard error.

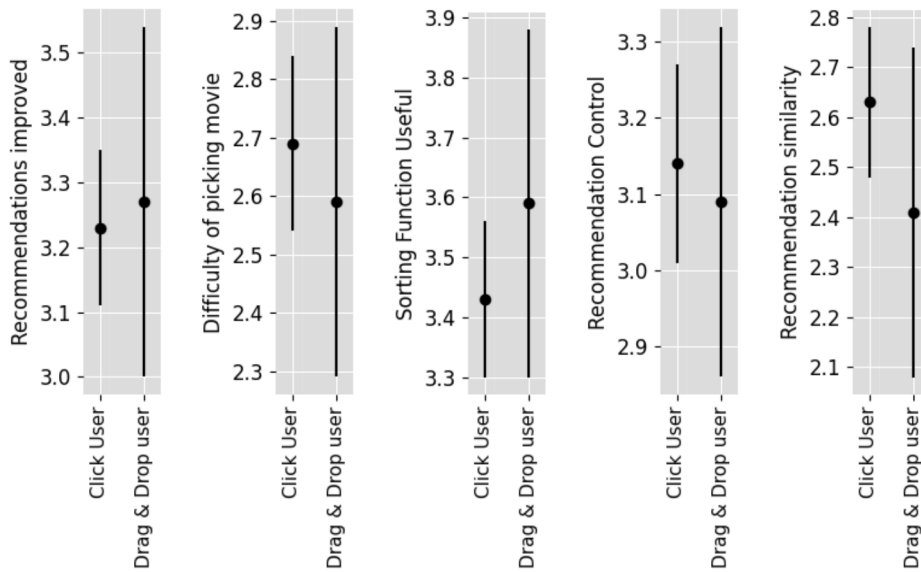


Figure B.9: System satisfaction metrics used for RQ2 in Section 4.2. Comparison plot with mean values and standard error.

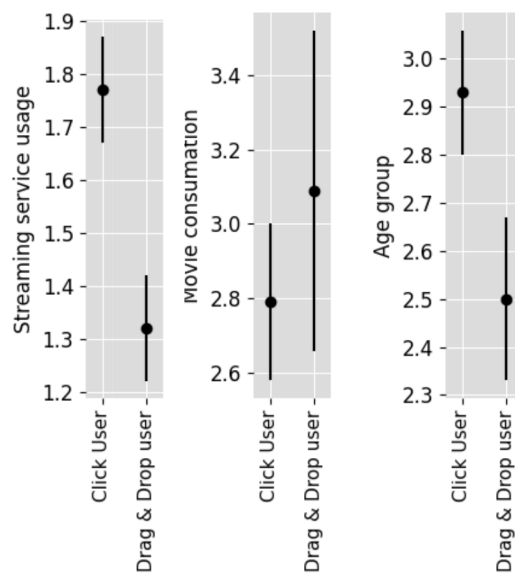


Figure B.10: Demographic metrics used for RQ2 in Section 4.2. Comparison plot with mean values and standard error.

## **B.3 RQ3**

### **B.3.1 Errorbar Plots for RQ3**

This section contains mean and standard error for list order satisfaction comparison for RQ3, included in appendix as alternative view of the values in [Figure 4.1](#).

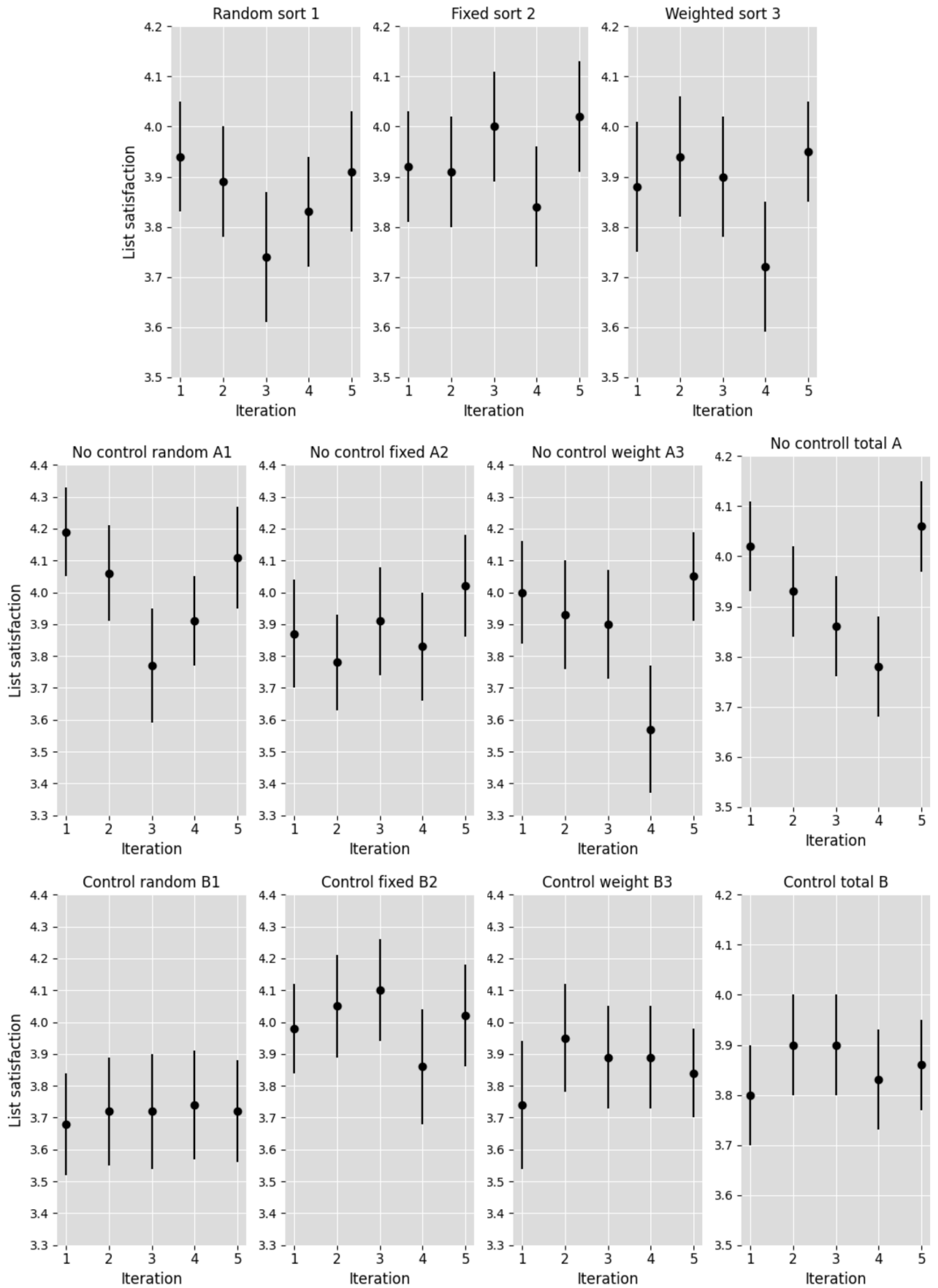


Figure B.11: List order satisfaction values alternative plot for RQ3 in Section 4.3. Comparison plot with mean values and standard error.

## **B.4 RQ4**

### **B.4.1 Demographic Variances Analysis Results RQ4**

These are additional analysis performed on demographic variables not included in the result chapter based on relevance.



Table B.6: Pairwise comparison with t-test between the genders of the participants. Means are displayed with standard error and are underlined in metrics where objective better values are present. N denotes sample size. Note: † = significant metric, \* $p < .05$ , \*\* $p < .01$ , \*\*\* $p < .001$ .

	$N_{Female} = 121$		$N_{Male} = 138$	
<i>System Usability</i>	Mean <sub>Female</sub>	Mean <sub>Male</sub>	$p$	Cohen's $d$
SUS Score	67.8±1.99	<u>71.1±1.49</u>	.178	-.17
Wish to use	3.45±0.11	<u>3.68±0.10</u>	.119	-.19
Complexity	2.42±0.12	<u>2.28±0.10</u>	.358	.11
Easy to use	3.98±0.10	<u>4.09±0.08</u>	.405	-.10
Assistance need	<u>1.85±0.12</u>	1.92±0.11	.669	-.05
Functions integration	3.59±0.10	<u>3.67±0.08</u>	.529	-.08
Inconsistencies	2.67±0.13	<u>2.47±0.11</u>	.233	.15
Easy to learn†	3.85±0.10	<u>4.1±0.07</u>	<b>.040*</b>	-.26
Cumbersome	2.51±0.13	<u>2.36±0.10</u>	.342	.12
Confidence using	3.83±0.10	<u>4.04±0.08</u>	.098	-.21
A lot to learn	2.14±0.12	<u>2.12±0.11</u>	.880	.02
<i>Participant Activity</i>				
Total list ordering	3.12±0.59	3.28±0.45	.828	-.03
Click list ordering	2.19±0.44	2.31±0.36	.834	-.03
Drag&Drop list ordering	0.92±0.30	0.97±0.24	.915	-.01
Times click was used	2.19±0.44	2.31±0.36	.834	-.03
Times drag&drop was used	0.27±0.09	0.37±0.09	.409	-.10
Movie index in list	1.67±0.07	1.77±0.07	.321	-.12
Selected list index	1.83±0.16	1.87±0.17	.868	-.02
<i>Recommendation Quality</i>				
List order satisfaction	3.88±0.08	<u>3.91±0.06</u>	.769	-.04
Diversity	3.83±0.09	<u>3.84±0.08</u>	.959	-.01
Novelty	3.80±0.11	<u>3.83±0.09</u>	.815	-.03
Serendipity	<u>3.66±0.10</u>	3.59±0.09	.607	.06
Relevance	3.3±0.11	<u>3.36±0.10</u>	.661	-.05
<i>System Satisfaction</i>				
Recommendation sim	<u>2.64±0.12</u>	2.33±0.11	.051	.24
Difficult selecting movie†	2.9±0.12	<u>2.56±0.11</u>	<b>.035*</b>	.26
Recommendation Improved	3.25±0.10	<u>3.39±0.10</u>	.309	-.13
Felt in control	3.25±0.12	<u>3.39±0.09</u>	.330	-.12
<i>Demographic</i>				
Sorting function useful	3.42±0.11	<u>3.56±0.09</u>	.336	-.12
Age category	2.81±0.09	2.83±0.09	.900	-.02
Streaming service usage	1.60±0.06	1.72±0.06	.800	-.18
Movies per week	2.92±0.17	3.10±0.16	.433	-.01

Table B.7: One-Way ANOVA on movies watched per week and age category of participants. Means are displayed with standard error and are underlined in metrics where objective better values are present. Note: † = significant metric, \* $p < .05$ , \*\* $p < .01$ , \*\*\* $p < .001$ .

<i>System Usability</i>	Movies watched per week		Age category of participant	
	$p$	$\omega^2$	$p$	$\omega^2$
SUS Score	.885	-.01	.174	.01
Wish to use*	<b>.001**</b>	.04	.209	.00
Complexity	.812	-.01	.071	.01
Easy to use	.131	.01	.652	.00
Assistance need*	<b>.005**</b>	.03	<b>.030*</b>	.02
Functions integration	.493	.00	.458	.00
Inconsistencies	.762	-.01	.247	.00
Easy to learn	.422	.00	.385	.00
Cumbersome	.729	-.01	.457	.00
Confidence using	.274	.00	.800	-.01
A lot to learn	.094	.01	.231	.00
<i>Participant Activity</i>				
Total list ordering	.423	.00	.478	.00
Click list ordering	.248	.00	.064	.01
Drag&Drop list ordering	.537	.00	.566	.00
Times click was used	.248	.00	.064	.01
Times drag&drop was used	.578	.00	.808	-.01
Movie index in list	.492	.00	.447	.00
Selected list index	.813	-.01	.098	.01
Preferred Control Element	.595	.00	.305	.00
<i>Recommendation Quality</i>				
List order satisfaction*	<b>.003**</b>	.04	.794	-.01
Diversity	.697	.00	.493	.00
Novelty	.341	.00	.423	.00
Serendipity*	<b>.041*</b>	.02	.110	.01
Relevance*	<b>.037*</b>	.02	<b>.012*</b>	.03
<i>System Satisfaction</i>				
Recommendation sim*	.888	-.01	<b>.036*</b>	.02
Difficult selecting movie	.816	-.01	.114	.01
Recommendation Improved*	<b>.013*</b>	.03	.339	.00
Felt in control*	<b>.004**</b>	.03	<b>.018*</b>	.02
Sorting function useful*	<b>.035*</b>	.02	.260	.00
<i>Demographic</i>				
Streaming service usage*	<b>&gt;.001***</b>	.20	<b>.023*</b>	.02
Movies per week	N/A	N/A	.390	.00
Age category	.543	.00	N/A	N/A
Gender	.193	.00	.911	-.01