École Polytechnique Fédérale de Lausanne

Investigation of User Perception on Fairness Issues in Two-sided Recommender Systems

by Chenkai Wang

Master Thesis

Approved by the Examining Committee:

Dr. Pearl Pu Faltings
Thesis Advisor

Dr. Li Chen
External Expert

Kalpani Anuradha Welivita; Ekaterina Svikhnushina
Thesis Supervisor

EPFL IC HCI Group
BC 145 (Bâtiment BC)
Station 14
CH-1015 Lausanne

February 25, 2023

Acknowledgments

I would first like to thank my thesis supervisor Dr. Pearl Pu of the Human Computer Interaction Group. She gave me precious suggestions and steered me in the right direction. I'm more than grateful that I have followed her to conduct my semester project from September 2021. She led me to understand the unnoticed glacier of user perceptions and fairness issues under the sea level.

In addition, I have to thank both my second supervisors, Kalpani Anuradha Welivita and Ekaterina Svikhnushina. Anuradha sent me dedicated assistance during the data collection parse. Ekaterina taught me to implement the survey questionnaire, and spared great effort to help me build the SEM model. Without their kind help, I would say it must be a thorny road filled with difficulties. I genuinely thank them for their support and understanding over these months. I would also like to acknowledge Dr. Yubo Xie for being a friend and colleague. And I am grateful for working with him last year.

Finally, I express my profound gratitude to my parents for providing me with unfailing support and continuous encouragement throughout my years of study. This accomplishment would not have been possible without them. I appreciate it from the bottom of my heart and wish them happy and healthy lives.

Lausanne, February 25, 2023 Chenkai Wang

Abstract

The fairness issue has been an urgent topic, with the popularity of mobile applications equipped with recommender systems. In this paper, we investigated the fairness problems from user's point of view through a survey. We prudently designed the survey questions and collected 1,300 responses in total. After filtering, 630 samples composed our *Dec2Dataset*. (Since the dataset is finally confirmed in December.) We embraced the common constructs of Perceived Ease Of Use and Perceived Usefulness. Meanwhile, we introduced the concept of Perceived Fairness to explore the relationships among these three principal constructs. We applied structural equation modeling methods to validate the model structure from collected data, proving our hypotheses' consistency. Through analysis with statistical reliability and validity, we confirmed that (1) Perceived Ease Of Use has a significantly positive influence on Perceived Fairness; (2) Perceived Fairness positively influences Perceived Usefulness; (3) these three constructs have both direct and indirect positive effects on Behavioral Intentions. This could bring a new perspective to researchers about the fairness problems in two-sided recommender systems.

Contents

Chapter 1

Introduction

Recommender systems (RecSys) have been applied in various scenes to help users find personalized, relevant items or products. Nowadays, RecSys are everywhere, including product recommendations in e-commerce systems, friend recommendations in social media, and hotel recommendations in trip planning platforms. With the favor of RecSys, these companies can explore and maximize their business interest. More precisely, online platforms, such as Netflix, Apple Music, and TikTok, are employing RecSys to help their customers find the most suitable product among enormous variety of products. Nevertheless, RecSys are never just human-made tools for seeking information. Indeed, the algorithms take charge of the allocations among different parties, such as video exposures to consumers, job opportunities to candidates. Further, a limited distribution can lead to unfairness and profit losses.

Artificial intelligence (AI) expands in almost every detailed aspect of today's life. AI has been shown as a "black box" from users' perceptions, potentially resulting in the lack of opacity and unfair discrimination. This draws attention from scientists to take care of potential ethical problems in areas related to RecSys. As revealed by Trustworthy AI framework [1], a famous multidimensional AI framework to help organizations bridge the ethics gap, the importance of addressing ethical challenges tends to be urgent. Today, businesses are rapidly expanding the scale and scope of their RecSys algorithms. It is hard to doubt that this trend has been aroused and accelerated by the COVID-19 pandemic, which is asking the majority to keep social distance and work from home. However, it also exposes these companies to threaten individual privacy and autonomy in unethical and inappropriate ways. While RecSys algorithms are increasingly taking control of the feeding content, the risks and challenges related to fairness issues have become even more critical and complex.

Like our human counterparts, RecSys are expected to adhere to social norms and ethics. People hope that RecSys make fair decisions in an explainable, consistent, transparent, and, most importantly, unbiased way. At the same time, figuring out the standard line of this appropriate method is not a piece of cake - even for ourselves. Generally speaking, RecSys learn their prospect from the training dataset. It is basically impossible to obtain datasets without any biases. Without proper care, RecSys will increase ethical problems exponentially. At the same time, this could amplify and propagate the biases. For example, a recommender system that decides where to place online job advertisements could tend to distribute male web browsers with well-paid job advertisements s more often. Because historical data shows that men generally earn more than women, resulting in typical and classical discrimination.

In a common situation, the majority of companies are dealing with a two-sided market with two kinds of stakeholders [2]: the first one is categorized as producers of goods and providers of services, which includes movie producers on Netflix, artists on Apple Music, and content providers on TikTok; the other one is customers and platform users. Traditionally, these business platforms only focused on maximizing customer satisfaction to gain the most profit on the users' end, where they lacked attention to producers and largely ignored the providers' interest during this process. Several recent studies have shown that a customer-oriented approach could undermine the healthy development of producers [3–6]. Since, in a multi-sided platform or two-sided platform, exposures to users have the most significant influence on producers' income. For instance, users, who have the chance to listen to a certain song, could help artists earn advertisement revenues. Additionally, a fixed percentage of users could buy digital albums or concert tickets. This principle of exposure also applies to TikTok. The exposure being monopolized by a few famous producers could lead to difficulties for new or lesser-known producers on the platform, potentially causing them to quit or seek alternative platforms, according to papers [5–7]. Without an illustration, this can dramatically damage users' experience and platform development. In this paper, we use TikTok as an example to explore the fairness issue in a two-sided platform.

TikTok, called Douyin in China Mainland, was first designed and launched in China in September 2016. Since the short-video platforms were planted a few years ago, Douyin took this advantage and quickly went viral in China. After that, its parent company, ByteDance, aimed to explore the foreign market and launched an international version of Douyin in the following year. In the beginning, TikTok mainly focused on lipsyncing and dancing videos. With the selection of the market, TikTok has developed and integrated into a multi-level, multi-field, and multi-business video recommendation platform. TikTok, as a recommender system with enormous users, has unparalleled influence and academic research value, which makes itself a suitable instance for us in this paper.

Although plenty of businesses are having a hard time during this post-pandemic era, TikTok is still growing steadily. As published in ByteDance's advertising resources, the latest statistics reveal that TikTok's advertising business grew faster in the first quarter of 2022 compared to the final quarter of 2021. The United States held the distinction of having the largest TikTok audience, with approximately 136.5 million users engaging with the platform, until April 2022 [8]. Currently, advertising investors can reach 970 million American users above 18 with advertisements on TikTok. The age and gender distributions among both USA TikTok users and the whole population are shown in Figure 1.1, whose data comes from resources [9, 10].

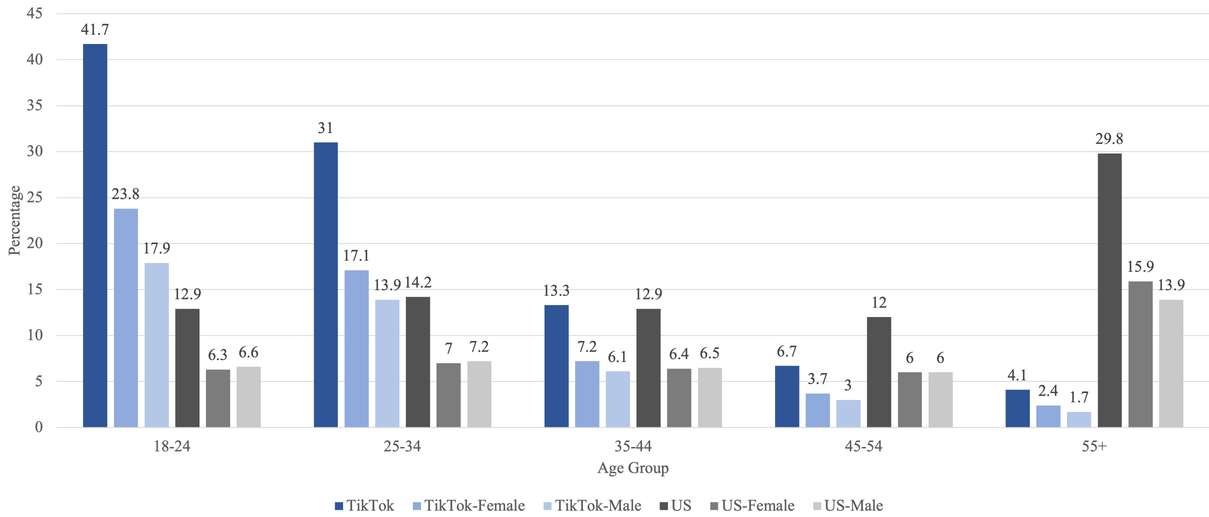

Figure 1.1: Age and gender distributions among both USA TikTok users and USA population.

This figure compares the age and gender distributions between the USA TikTok users and USA overall population. In each age group, from left to right, these six bars respectively represent:

- the percentage of the USA TikTok users to the whole USA TikTok users;
- the percentage of female USA TikTok users to the whole USA TikTok users;
- the percentage of male USA TikTok users to the whole TikTok users;
- the percentage of the USA population to the whole USA population;
- the percentage of the female USA population to the whole USA population;
- the percentage of the male USA population to the whole population.

With data from April 2022, the figure from April 2022 displays that 24% of TikTok's domestic audience consisted of women aged 18 to 24 years old. At the same time, male users in the same age group made up approximately 18 percent. Approximately 17% of female users aged 25 to 34 years old are active on TikTok, and 14% of male TikTok users in the USA belong to the same age group. From this figure, we can easily find that the audience of TikTok in the United States congregates of younger generations. Additionally, in each age group, female users are always more than male users, concerning the changing female-male ratio with the change of different age groups.

Chapter 2

Backgrounds & Related Works

2.1 Fairness Issue in Recommender Systems

Why is the fairness issue a serious problem in RecSys? An intuitive example can illustrate the importance here. With limited recommendation slot positions on Amazon, the question arises as to which producers' goods should receive the precious exposure opportunity to users. Predictions using data appears to be a suitable answer to this question, however, bias may arise as most RecSys models are trained with machine learning (ML) models based on a specific training dataset. The training data in this process may carry social bias, which may then be learned by recommendation models as alternative timesavers, thereby echoing or even exacerbating the bias in the training data. Once the unfairness is built into the algorithms, information asymmetry is impossible for certain users with discriminated properties. This phenomenon can result in aggravation with the Matthew effect [11].

From the standpoint of platform benefit, to ensure both compliances with legal regulations and the long-term sustainability of the platform, the executors should take care of the fairness issues. Naturally, famous accounts (for example, well-known film actors) generally receive more exposure for their videos, while videos created by grassroots may be unnoticed and ignored. The struggling dilemma may cause these users to leave the platform, who are also the majority source of benefit for the platform. The decreased diversity and activity of the platform's content will negatively impact its sustainability. These above instances are only part of the potential harm to the platform's development. Recently, Researchers have proposed various definitions of fairness. Sühr et al. [5] discuss the fairness definition in ride-hailing with the example of Uber. Meanwhile, Patro et al. [4] introduce related concepts using the e-commerce method on Amazon. These definitions focused on different aspects of the fairness issue in macro machine learning problems. Here we will go through some of the fairness definitions that are most relevant to our study.

To achieve group fairness, the treatment of the protected groups should be comparable to that of the advantaged group, such as the different demographic groups represented in a dataset. Group fairness is paid attention to specific situations, like inclusive hiring for women groups or recommendations for inactive groups. It is sharp in TikTok that there is a noticeable disparity in recommendation quality among users with heterogeneous activeness, such as the number of interactions. The COMPAS controversy [12] emphasized the importance of evaluating group fairness using model properties such as accuracy and false-positive rate, which are two of the most prominent examples [13]. Additionally, there are other works related to group fairness. Fu et al. [14] address the issue of group unfairness in explainable recommendation over knowledge graphs using a fairness-constrained approach. Articles [3, 15] categorize different types of multi-stakeholder platforms and further discuss various group fairness properties.

In contrast, the principle of individual fairness dictates that comparable individuals should receive comparable treatment, which is emphasized in problems related to recidivism prediction and loan decisions. Cynthia Dwork et al. [16] first introduced the concept in the paper called "Fairness Through Awareness" in 2012, which was one of the seminal and foundational works in the fairness-related area. Individual fairness, as opposed to group fairness, imposes restrictions on the treatment of each pair of individuals, making it a more specific concept. Patro et al. [4] consider individual fairness from the perspective of both producers and customers. And the researchers respond to the question of the long-term sustainability of two-sided platforms. Some interesting links between the two concepts are (1) Based on Dwork et al. [16], group fairness does not guarantee individual fairness or vice versa. (2) Based on Biega et al. [17], individual fairness, under certain circumstances, can promote group fairness.

User fairness considers fairness based on users' perspectives, which means RecSys should recommend the same product query to users with the same demographic attributes. Abdollahpouri et al. [18] examine the fairness issue from the perspective of users, investigating how popularity bias leads to deviations in recommendations from users' expectations on the RecSys. As for item fairness, RecSys requirements related to fairness issues may come from items' perspectives, which could also mean the perspectives of products and producers. For instance, a search for "phone case" on Amazon might result in accessories for the iPhone appearing at the top of the list, while those for other brands receive lesser visibility. The lack of other brands' accessories is an item-side unfairness for them. For item fairness, there are many studies concerning the potential damage from popularity bias in RecSys. According to research [19–21], increasing the number of long-tail items or the overall catalog coverage is often a feasible solution to the problem.

2.2 Envy-freeness in two-sided recommender systems

The fairness of RecSys is studied mostly from either the perspective of customers (User Fairness) or product/services providers (Item Fairness) in most studies. And these studies are conducted exclusively from one perspective, which are also known as single-sided fairness. At the same time, multi-sided fairness, which considers both customer-side and provider-side fairness, is another kind of fairness. With experiences from companies that survived in the information era, the sustainable development of RecSys platforms needs evolutions on both the product and customer sides. Guaranteeing fairness for one group might sacrifice the fairness and benefit of the other one [22, 23]. With an increasing awareness of the provider-side importance, studies concerning the two-sided fairness [4, 5, 23], and the multi-sided fairness [3, 15, 22] have drawn much more attention. Generally, the two-sided objective is a linear interpolation of consumer and producer fairness metrics [5, 22]. Its general form is:

$$\lambda \times Inequality_{product\ side} + (1 - \lambda) \times Inequality_{customer\ side}$$

Several specific fairness concepts are incompatible or mutually exclusive, which is unfortunate. Both Kleinberg et al. [24] and Chouldechova et al. [13] claim that equal accuracy in risk scores and balance in risk quantiles across groups can only be attained under specific and stringent conditions.. Grgić-Hlača et al. [25] argue that achieving both process fairness and outcome fairness sometimes may come at the cost of accuracy. As said, Realistic ML models often cannot meet all desired definitions of fairness simultaneously due to the mutual exclusivity of many fairness definitions. Considering these inherent trade-offs, we acknowledge that the model design could satisfy one definition of fairness while violating another.

Therefore, to address this issue, Envy-freeness provides a feasible solution. Envy-freeness was initially studied in fair allocation from Foley's study in 1966 [26]. More recently, fair classification [27, 28] stipulates that it is considered fair to apply different strategies to different people as long as it is benefiical to every individual. Based on this theory, we consider the RecSys as fair only if it distributes better content according to individuals' preferences. In contrast, we consider it unfair not to give users a better recommended content when there is such a solution existing.

Envy-freeness aligns with providing users with their most enjoyed recommendations, which is exactly different from parity or equal utility. The sources of envy can be diverse. Compared to equal user utility, an improvement of envy-freeness is that it involves interpersonal comparisons.

Chapter 3

Methodology

3.1 Survey Design

We developed an online questionnaire and hosted it on the EPFL server to validate our conceptual model and hypotheses, mentioned in Section 3.2. This can avoid the potential leakage to third-party survey agent platforms. The survey aimed to understand participants' interaction experiences with TikTok and some issues arising from this process. Based on their recent TikTok usage, we asked participants to evaluate their experience on the following themes: system quality, service quality, information quality, perceived fairness, perceived ease of use, perceived usefulness, attitude towards use, and behavioral intentions. These themes also function as the principal constructs in our later SEM model analysis. Based on a validated scale aiming to measure Web quality and playfulness [29], we designed our survey questionnaire and modified other questions to adapt to our situation. It is worth noting that these themes are not disclosed to the participants. They would only receive a bunch of questions and reply to these questions. The participants did not act based on the conceptual themes.

The survey website consisted of five web pages: information sheet, consent form, demographic questions, main questions, and end page. The questionnaire covered six demographic questions and 45 thematic questions, which can be found in Appendix A.1. Demographic questions included gender, age, English fluency, profession, usage length, and weekly usage duration. The 45 main questions are 5-point Likert-type items, which is a typical psychometric response scale where responders specify their level of agreement in five magnitudes: from "Strongly disagree" (1) to "Strongly agree" (5). The questions adapting the 5-Likert-type scale can be seen in Appendix A.1. The number in the parentheses is the score we transfer categorical features into numerical ones during the data processing stage.

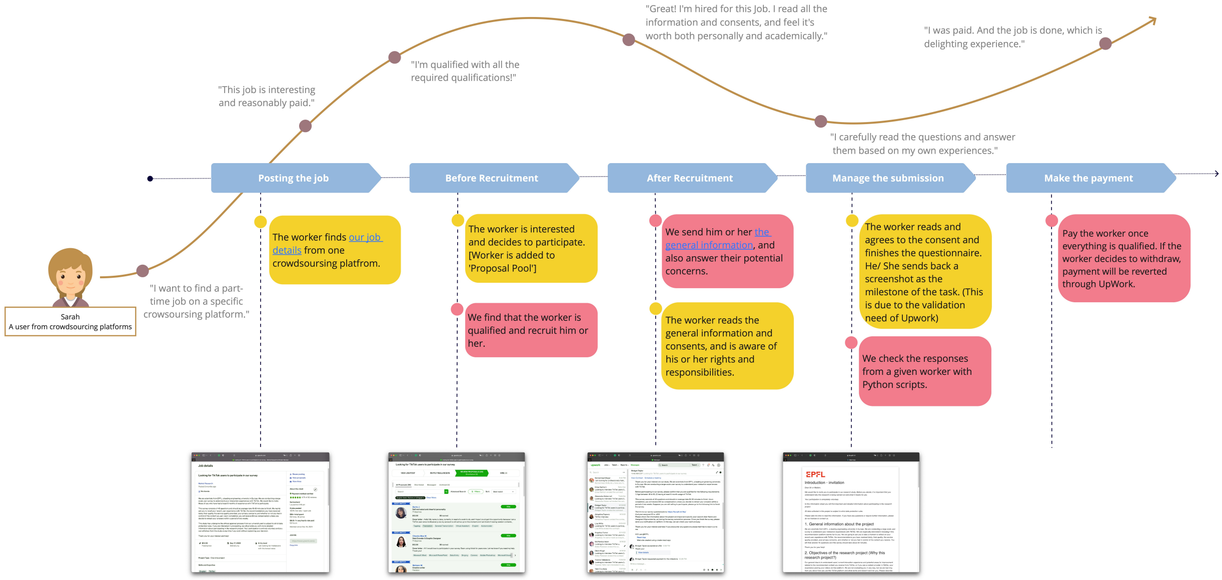

Figure 3.1: User Journey Diagram for our TikTok Survey.

A user journey diagram usually refers to a visual trip across the possible solution. In the user journey diagram, we need to consider not only the user's steps but also their feelings, pain points, and possible difficulties along the road. The user journey diagram of participating in our TikTok survey is in Figure 3.1. The process of an example user, Sarah, participating in our survey is presented. The brown line is her inner mental activity, as we speculated. The blue line is the task we addressed from our perspective. And all red and yellow bubbles are the interactions between the participant and us. This significantly helps us to pay attention to users' perceptions during the survey.

3.1.1 Attention Checks

Reflected by plenty of studies, many MTurker responses could not be usable due to high rates of MTurk workers' ignorance and inattention. Approximately 10% of workers fail attention check questions in studies using Mechanical Turk, which is a typical finding from the article [30]. Including these data will damage the quality of our dataset and may cause either false positives or false negatives during the analysis step [31]. Therefore, it is undoubted and warranted that we proceed removal of such data. To cope with this dilemma, incorporating attention-check questions in our Mechanical Turk research is justifiable and reasonable in order to detect and eliminate non-compliant responses. In our study, we designed seven reversed question pairs to check MTurk workers' attention, covering six principal constructs, which significantly mitigate drawbacks related to data quality.

- It's easy to navigate on TikTok to find what I want.
  TikTok has unclear app navigation.
- The recommended items correspond to my needs and preferences.
  TikTok wrongly infers my interests.
- The recommended items cover a variety of topics.
  TikTok recommended items are repetitive.
- It is easy for me to become skillful at using TikTok.
  Using TikTok requires a lot of mental effort.
- TikTok gives me enjoyment and keeps me entertained.
  Using TikTok is dull and boring.
- Using TikTok is a good idea.
  Using TikTok is a bad idea.
- I will use TikTok on a regular basis in the future.
  I don't intend to use TikTok in the future.

In the above, these reversed question pairs can also be found in Appendix A.1. Specifically speaking, the standard of correction in a reversed question pair is evaluated below. A reversed question pair, called Question A and Question B here, is considered correct if Question A gets a score of [1,2,3] and Question B gets a score of [3,4,5]. To encourage participants to pay attention to the questionnaire, we declared that there are several attention checkers in our survey. And if the participant got most of the attention checkers correct, we would pay a bonus reward. In practice, a bonus reward of 20 cents was distributed to the worker if he/she got correct in at least 5 of 7 question pairs.

3.2 Model Development

This paper established the hypothesized model from the original Technology Acceptance Model (TAM) [32]. TAM initially aimed to explore and analyze the users' acceptance of information technologies (IT). Earlier, TAM was generated from the Theory of Reasoned Action (TRA) created by Fishbein and Ajzen [33]. This theory posits that a person's disposition towards performing a particular action drives their social behavior. According to one's beliefs, the reason for conducting a specific behavior is that the advantages of performing this behavior should outweigh the disadvantages, which is a consequence after evaluating the value of those outcomes. The Theory of Reasoned Action (TRA) asserts that a person's behavior is directly influenced by their deliberate intention to perform, as individuals generally act in accordance with their intent.

Adopting TRA's causal links, TAM explained the individual acceptance behaviors of IT tools on the personal layer. It suggested that perceived usefulness and perceived ease of use are two principal determinants during the usage of an IT-related application. Consistent with TRA, users' beliefs determine their attitudes toward using the application in the TAM model. Behavioral intentions to use, in turn, are determined by their attitudes toward using the application. Finally, behavioral intentions to use can quickly lead to actual use shortly in the future.

Previous research has tested and validated this model's utility through various experiments. Additionally, in 2001, Moon and Kim extended the TAM model for the WWW context [34]. The study encompassed three Web elements for e-commerce: system quality, information quality, and service quality. In 2007, Ahn et al. [29] examined the impact of playfulness on users' adoption of e-commerce. They assessed the correlation between Web quality factors and the user's behavioral intentions. Here, we embraced these ideas and developed them to adjust to the current popularity of mobile applications.

Furthermore, we proposed the following principal constructs. These constructs were derived from previous research, which established and validated users' perceptions of mobile applications. We selected and summarized the most commonly used items: "Perceived Ease of Use", "Perceived Usefulness", "Attitude Toward Use", and "Behavioral Intentions to Use" as well. These constructs demonstrated substantial reliability and internal consistency in Chapter 5. At the end of this chapter, we formed a set of hypotheses regarding the relationships among the constructs To evaluate our model.

3.2.1 Concerns

In the principal construct of "Concerns", we use exploratory factor analysis(EFA) to reorganize system quality and service and information quality. In our case, EFA was used to reduce data to a smaller amount of variables and explore the underlying theoretical structure behind the observed phenomenon. Based on the factor loading result, we successfully extracted the following four items to compose the "Concerns" construct.

- "TikTok has unclear app navigation."
- "TikTok wrongly infers my interests."
- "TikTok recommended items are repetitive."
- "TikTok displays too many advertisements."

We hypothesized that "Concerns" aggregates users' concerns about the TikTok platform, which negatively affects "Perceived Fairness". From a macro perspective, these concerns can also decrease users' intentions to use TikTok in the future.

3.2.2 Service and Information Quality

Service quality and information quality were combined in our proposal because TikTok, as a recommender system, provides services of recommending videos, which provides information to users simultaneously. From this standpoint, we further divided "Service and Information Quality" (SerInfQ) into six subconcepts: quality, accuracy, diversity, agency, adaptiveness, and inferred privacy.

Quality refers to the quality of recommended videos by the platform. Accuracy can be separated into personalization width and personalization depth. Personalization width means the comprehensiveness of recommended content, while personalization depth focuses on users' more niche hobbies. Diversity is evaluated by the question item "The recommended items cover a variety of topics.", which is quite tricky and obscure for the platform. People can hear complaints that the recommended content is too diverse in one area and too identical in another. In users' perception, the desired diversity is satisfied with an expectation of diversity. This criterion can be seen as serendipity as well. Agency concerns users' ability to control their feed and edit their personal preferences. As for adaptiveness, it emphasizes the adaptability of recommended content. Called personalization sensitivity, it also focuses that while a person's life stage changes, the recommended content should evolve and adapt to the user's needs. In inferred privacy, the leading indicator is whether users' provided personal information is adequately protected.

In summary, we hypothesized that "Service and Information Quality" positively affects users' "Behavioral Intentions". This hypothesis can be broken into that: "Service and Information Quality" has a positive effect on "Perceived Ease of Use" and "Perceived Usefulness".

3.2.3 Perceived Fairness

In general, bias refers to an unfavorable prejudice towards an individual or group, often considered unfair. We addressed "Perceived Fairness" (PF) to correspond to five biases extracted from the general fairness issue in recommender systems. In this paper, we covered demographic bias, user interaction bias, social connection bias, popularity bias, and envy-freeness.

Generated due to inferred or disclosed demographic information in the TikTok platform, demographic bias could be the most common unfairness. User interaction bias (feedback bias) happens when the users' engagement with the output of RecSys influences subsequent training data. For example, user interactions were used as new input to refine the recommendation algorithm over time, thus creating a feedback loop. In our case, this generates from a user interacting with the TikTok platform: the user can tap on the heart button to like the recommended video, select the speech bubble to comment, and click the forward icon to share content. Part of the user interaction bias may also come from searching in TikTok. Observation bias is also included in the concept of user interaction bias. Observation bias emerges when the system recommends content similar to what the user previously rated high, creating the filter bubble phenomenon [35]. Popularity bias (item popularity bias) suggests that the most popular items will appear higher in search, further attracting more engagement. Social bias occurs when the actions of others impact our judgment. For instance, when content consumers give low ratings or negative reviews to an item, they could tend to align with high-rating comments. In other words, they adjust their object evaluation to be consistent with the majority. Specifically in TikTok, when we find some videos not so interesting, yet these videos received many views and likes, we may stay longer thinking that perhaps we have not got the point. Last but not least, envy-freeness means the RecSys are reckoned fair if each user prefers their own recommended content compared to others [36]. Envy-freeness allows a recommender system to be fair even if there are disparities between groups as long as the recommended content satisfies user preferences. Similar studies have been conducted in other RecSys tasks under the shelter of preference-based fairness problem [28, 37, 38]. Recommendations based on envy-freeness are the proper extension of these approaches to personalized RecSys.

In summary, we hypothesized that "Perceived Fairness" positively affects "Attitude Towards Use". Further, it will positively affect "Behavioral Intentions". Also, we hypothesized that "Perceived Ease of Use" has a positive effect on "Perceived Fairness". Then "Perceived Fairness" positively affects "Perceived Usefulness". This structure forms the TAM loop in our design.

3.2.4 Perceived Ease of Use

Here are two fundamental variables among many variables influencing users' behavioral intentions. First, people tend to use an application based on their belief that its advantages outweigh the disadvantages of using this application. This first variable is referred to as "Perceived Usefulness". Additionally, even if potential users perceive the application to be useful and practical, they may also view the system as too complex or cumbersome to use. This opinion reveals that the effort of using the given application outweighs the benefits of usage. This second variable is theorized as "Perceived Ease of Use".

As quoted, "Perceived Ease of Use" (PEOU) is "the degree to which a person believes that using a particular system would be free of effort" [32]. This follows from the dictionary definition of ease: "freedom from difficulty or great effort." Effort can be a finite and limited resource that a person has to allocate across numerous tasks. With all other variables being identical, an application perceived as more straightforward to use than another is more likely to be accepted by users. Therefore, "Perceived Ease of Use" is hypothesized to have a significantly positive effect on "Behavioral Intentions".

3.2.5 Perceived Usefulness

Meanwhile, Davis defined "Perceived Usefulness"(PF) as "the degree to which a person believes that using a particular system will improve his or her job performance" [32]. As commonly believed, "Perceived Usefulness" plays a crucial role in determining the acceptance of future use. This concept follows from the definition of the word useful: "capable of being used advantageously". In David's study [32], the two-factor analysis suggested that "Perceived Usefulness" and "Perceived Ease of Use" are statistically distinct constructs. "Perceived Ease of Use" is hypothesized to affect "Behavioral Intentions significantly".

3.2.6 Attitude Towards Use

The construct "Attitude Towards Use" (Att) refers to an individual's feelings and dispositions involving using an application. Gibson et al. [39] defined attitude as a "positive or negative feeling or mental state of readiness, learned and organized through experience, that exerts specific influences on a person's response to people, object, and situation". In 1971, Triandis [40] claimed that attitude should consist of affective, cognitive, and behavioral components. The affective component of attitude is the emotion or feeling, which includes statements of likes or dislikes about particular objects. The cognitive component of attitude is statements of beliefs. In other words, an individual believes that a particular object can significantly increase the quality of her or his output. The behavioral component of attitude is what an individual does or intends to do and is affected by individuals' experience [41]. We hypothesized that "Attitude Towards Use" has a strong positive relationship with "Behavioral Intentions".

3.2.7 Behavioral Intentions

"Behavioral Intentions" (BI) generally refers to users' decision to use this application in the future. According to the Theory of Planned Behavior [42], BI has been shown to predict actual behavior effectively. Therefore, evaluating the BI of TikTok users based on their experience is valuable for gaining insight. As done in previous studies related to users' adoption [43, 44], our evaluation of the "Behavioral Intentions" construct involves assessing the users' intention to continue using TikTok and their desire to recommend it to others.

3.3 Model Hypotheses

After giving the above definitions, we organize a set of hypotheses about possible relationships among these principal constructs, which can be found in Figure 3.2. In the following paragraphs, we shorten "Service Information Quality" to SerInfQ, "Perceived Ease of Use" to PEOU, "Perceived Fairness" to PF, "Perceived Usefulness" to PU, "Attitude Towards Use" to Att, and "Behavioral Intentions" to BI. First, we hypothesize that there is only one negative correlation arrow from "Concerns" to PF. Since concerns present users' harmful concerns about the TikTok platform, the more concerns, the more unfair the platform is reckoned. Then SerInfQ has a positive influence on both PEOU and PU, which is consistent with our intuition. PEOU affects PU through PF. Further, PF can positively influence the Att. Lastly, Att, PEOU, and PU have positive correlations to BI. In the following chapters, we present the analysis and evaluation of our model hypotheses.

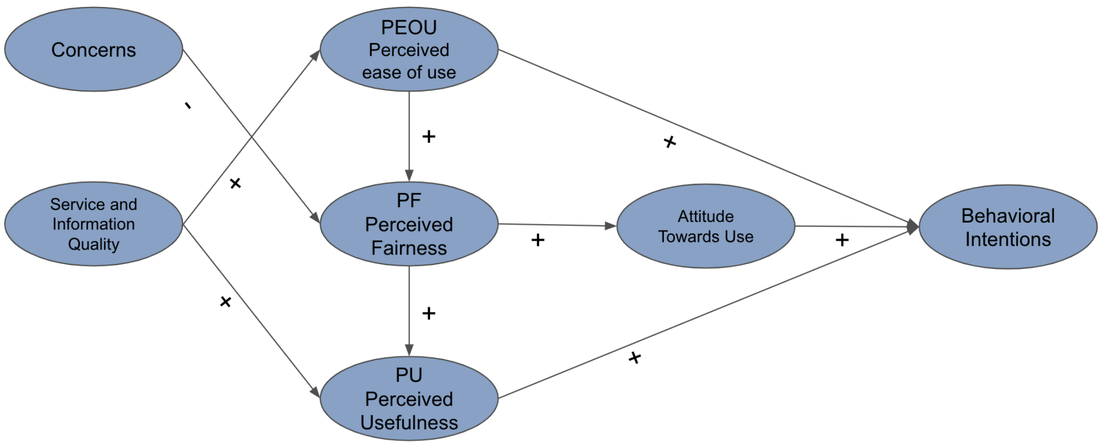

Figure 3.2: Model hypothesis.

Chapter 4

Dec2Dataset

4.1 Data Collection

In our research, we used Amazon's Mechanical Turk (MTurk) [45] to collect our data samples. MTurk is a crowdsourcing platform that allows users to post tasks and receive responses from workers. Its wild applicability enables requesters to publish assignments like survey participation, data validation, content moderation, etc.

According to a review article [46], the MTurk platform's popularity can be attributed to four benefits: (a) extensive and diverse participant pool, (b) ease of access and speed of data collection, (c) reasonable cost, and (d) flexibility regarding research design choice. The use of web-based research using MTurk has increased tenfold over just the last decade [47], making it the most frequently used online data collection method by far, as revealed by the study [48]. Furthermore, researchers [47, 49] have given positive support for samples drawn from commercial online panel data (OPD). A recent study [47] conducted meta-analyses of OPD, and then compared internal reliability estimates with conventionally sourced data. The authors suggested, with appropriate caution, that OPD is suitable for many exploratory research questions. With the above support, we chose to use the MTurk platform to conduct our study.

In the process of recruiting MTurk workers, applying qualifications is validated as the most effective and convenient strategy for improving the quality of collected Amazon MTurk datasets. Qualifications, acting as filters, allow researchers to select particular worker demographics features or working experiences with MTurk for specialized projects. This enabled us to distribute task assignments to particular groups of workers on MTurk. Our study applied the following qualifications. We required the HIT approval rate of AMT workers to be higher than 98%. Here, the HIT approval rate represents the proportion of completed Human Intelligence Tasks(HITs) that Requesters approve. To ensure the workers are familiar with the MTurk platform, we asked them to have more than 500 Approved HITs. We further asked all the participants to be residents of the United States, since we hoped to focus our study on the USA scope. Furthermore, we distributed a bonus reward of 20 cents to participants who successfully finished our attention check. The detailed criteria are in section 3.1.1.

We prudently chose our survey's sample to ensure that their experience with TikTok could truly reflect their user perceptions during the survey. We carefully evaluated the validity and reliability of the hypothesized model to address any limitations of the self-report data collection method. The details are provided in the subsequent sections.

First, we launched two pilot studies(N=20,200, respectively) in September 2022. The purpose of these two pilot studies is to verify the feasibility of our data collection process and evaluate the efficiency of survey questions. We modified our survey questionnaire based on participants' feedback and relaunched our formal studies in November 2022. Due to scientific rigor, we did not include the sample from pilot studies in our *Dec2dataset*.

Formally, we published 1,300 assignments through the MTurk platform. The survey was posted as 12 separate batches. As mentioned, the same qualifications were applied. The reward for participating in our survey was 1.82 dollars. We accomplished the whole study between November 15th and December 2nd, 2022.

4.2 Data Screening

Based on our two pilot studies, we developed the rules for data screening. We thoughtfully adopted feasible and reasonable approaches to exclude noisy submissions. In summary, the following criteria were implemented to filter our collected dataset:

- Filter 1: Exclude participants with more than 40 same answers.
- Filter 2: Exclude participants who choose the "Other" option in the language question. (The 3rd demographic question in Appendix A.1)
- Filter 3: Exclude participants who choose the "Other" option in the experience question. (The 5th demographic question in Appendix A.1)
- Filter 4: Exclude participants using less than 3 minutes.

To begin with, with 1300 assignments posted, 1287 workers completed and submitted the survey. Thirteen responses were missing as these 13 workers failed to submit a valid survey form. This could be caused by misusage while filling out the survey. There was no missing data in the sample because participants could only submit their responses if all the questions were completed.

The responses were collected in "JSON" format and converted into Pandas DataFrame. Applying Filter 1, we discarded 18 responses as these workers selected the same answer over 40 times in 45 main questions. Applying Filter 2, we excluded 0 workers with the "Other" option in the "Please specify your English fluency" question. This left us responses with options "Native" and "Fluent". Applying Filter 3, we excluded 266 workers with the "0-6 months of usage" or "Other" options in the "Please specify your TikTok experience" question. This left us responses with the option "Over 6 months of usage". Lastly, we evaluated each worker's elapsed time from opening the survey to the end to avoid impairment from automatic survey-completing bots. The submission was considered invalid if the elapsed time was less than 3 minutes. Because during the survey, the first step was to read the information sheet and consent form. Then 6 demographic questions and 45 main questions were followed. It was unlikely for a human with average reading and clicking speeds to answer every question in 4 seconds, despite reading all the descriptive paragraphs. We filtered out 373 such cases. In total, we discarded 670 responses with the above filtering criteria. Therefore, the final sample size of our *Dec2Dataset* became 630. According to the rule of thumb, having at least 10 participants for each survey question item[50], our sample size is adequate for a steady factor estimate analysis.

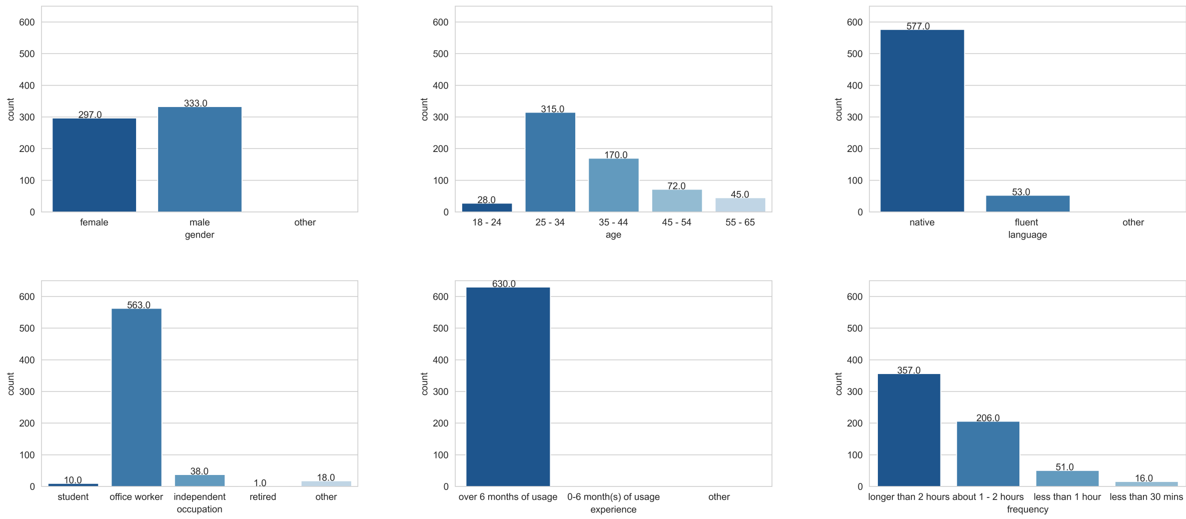

Figure 4.1: Demographics of *Dec2Dataset*(N=630).

4.3 Data Description

In total, 630 responses were gathered and prepared for our model generation. The distributions of 6 demographic features are shown in 4.1. Females(52.9%) and males(47.1%) were approximately equally distributed. 4.4% of participants were from the 18-24 age group, and 50% of them were within the 25-34 age group, 27% of them were from the 35-44 age group, the rest(18.5%) of

Demographics	Item	Frequency	Percentage
Total		630	100%
Gender	Male	333	52.9%
	Female	297	47.1%
Age	18-24	28	4.4%
	25-34	315	50.0%
	35-44	170	27.0%
	45-54	72	11.4%
	56-65	45	7.1%
language	Native	577	91.6%
	Fluent	53	8.4%
	Other	0	0%
Occupation	Student	10	1.6%
	Office worker	563	89.4%
	Independent	38	6.0%
	retired	1	0.2%
	other	18	2.9%
Experience	Over 6 months	630	100.0%
	0-6 month(s)	0	0%
	Other	0	0%
Frequency	Longer than 2 hours	357	56.7%
	About 1-2 hour(s)	206	32.7%
	Less than 1 hour	51	8.1%
	Less than 30 mins	16	2.5%

Table 4.1: Profile of participants(N=630).

them belonged to users above 45. As mentioned in 4.1, the survey was conducted in the United States, and the majority(91.6%) of participants were native English speakers. Meanwhile, the rest(8.4%) of them are fluent English speakers. According to the collected responses, 89.4% of the sample were office workers, which is in line with the approximate age distribution.

According to the section 3.1.1, we mentioned that a bonus reward was distributed to participants who got correct in at least 5 of 7 reversed question pairs. In total, 429 participants received this 20 cents bonus, which accounted for 68% of approved workers.

We also looked into their usage length and frequency with TikTok. After a thoughtful decision, we filtered out participants with less than 6 months of TikTok usage, indicating that the whole sample was lately active on TikTok and had more than 6 months of usage experience. Their current weekly usage was diverse: 56.7% of them were on TikTok longer than 2 hours each week;

32.7% were using TikTok for about 1-2 hour(s) weekly; 8.1% were using TikTok for less than 1 hour; and 2.5% were using TikTok less than 30 minutes. Therefore, we can safely conclude that our participants were familiar with and involved with the TikTok platform lately. The profile of 630 participants is summarized in Table 4.1.

	Skewness	Kurtosis
count	45.000000	45.000000
mean	-0.648472	0.281799
std	0.430661	0.835122
min	-1.251288	-1.321952
25%	-0.914921	-0.196441
50%	-0.718030	0.387247
75%	-0.603526	0.891446
max	0.631627	1.803923

Table 4.2: The statistics of skewness and Kurtosis for 45 main questions.

According to research [51], all the variables should be well modeled by a normal distribution to satisfy SEM assumptions. The distributions of each question can be found in Appendix A.1 and A.2. We also annotated each question's mean, standard deviation, and median values. As shown, the empirical distributions considerably deviated from the bell-curved distribution. Further, we checked the skewness and kurtosis of our 45 main questions. The skewness (-1.25, 0.63) and kurtosis (-1.32, 1.80) of all variables fall into the (-2.00, 2.00) interval, which is the recommended acceptable range of normality test for Likert-type questions[52]. Thus, our data also meet the normality requirements. The statistics of skewness and kurtosis for 45 main questions can be found in Table 4.2.

Chapter 5

Results

5.1 Model Validity and Reliability

We conduct the following data analysis and model evaluation using the Lavaan version 0.6.12. We considered the model validity and reliability from 3 dimensions: basic statistics, internal reliability, and convergent validity. Also, we utilized the maximum likelihood (ML) estimation with robust standard errors.

We discarded the question items with factor loading less than 0.5 to maximize the model interpretability. After this, we obtained 23 question items in this step of our model structure, presented in Table 5.1. To identify the reasonable clusters of question items, we evaluated internal reliability with Cronbach's alpha and item-total correlation. Both item-total correlation and Cronbach's Alpha range from 0 to 1. And higher values indicate a more significant internal consistency. Attention to internal reliability aimed to reveal the consistences of model measurement structure. Meanwhile, after carefully selecting the question items to build the principal constructs, we achieved that all seven principal constructs have a more than 0.6 Cronbach's alpha, which meets the standard of 0.6 for Cronbach's alpha [53]. Except for two principal constructs (PEOU, PU), Cronbach's alpha of the rest five principal constructs are more than 0.65. Furthermore, the highest Cronbach's alpha is 0.79, which is accomplished in PF with question items [Q19, Q20, Q21, Q22].

As for the item-total correlation in our research, the lowest value is 0.62, and the highest is 0.88. Generally, item-total correlation serves as a criterion of assessment and purification. There are several standards from different studies: 0.30 threshold of item-total correlation by Cristobal et al. [54], 0.40 threshold of item-total correlation by Loiacono et al. [55], 0.50 threshold of item-total correlation by Francis and White [56] and Kim and Stoel [57]. We went over the most

Construct	Item	Statistics			Internal reliability		Convergent validity		
		Mean	SD	Median	Cronbach's alpha (0.6)	Item-total correlation (0.5)	Factor loading (0.5)	Composite reliability (0.6)	Average variance extracted (0.5)
Concerns					0.69			0.70	0.47
	Q2	2.64	1.32	2		0.87	0.74		
	Q7	2.89	1.37	3		0.88	0.72		
Service & Information Quality					0.68			0.66	0.72
	Q4	3.84	0.93	4		0.72	0.51		
	Q6	4.01	0.83	4		0.64	0.53		
	Q8	4.09	0.78	4		0.62	0.51		
	Q10	3.90	0.95	4		0.67	0.54		
	Q13	4.03	0.85	4		0.67	0.56		
Perceived fairness					0.79			0.80	0.51
	Q19	3.72	1.06	4		0.78	0.68		
	Q20	3.70	1.05	4		0.83	0.79		
	Q21	3.79	0.99	4		0.79	0.73		
	Q22	3.66	1.05	4		0.74	0.61		
Perceived ease of use					0.61			0.61	0.66
	Q26	3.97	0.89	4		0.77	0.63		
	Q27	4.03	0.86	4		0.74	0.54		
	Q28	4.32	0.86	5		0.74	0.58		
Perceived usefulness					0.64			0.65	0.68
	Q24	4.17	0.82	4		0.69	0.52		
	Q30	4.25	0.79	4		0.73	0.60		
	Q33	4.16	0.79	4		0.63	0.54		
	Q34	4.13	0.80	4		0.72	0.59		
Attitude toward use					0.65			0.65	0.52
	Q36	4.17	0.87	4		0.86	0.71		
	Q38	4.16	0.85	4		0.86	0.68		
Behavioral intentions					0.65			0.59	0.68
	Q40	4.22	0.80	4		0.77	0.58		
	Q41	4.13	0.84	4		0.76	0.52		
	Q44	4.23	0.85	4		0.78	0.60		

Table 5.1: Summary of internal reliability and convergent validity.

rigid bar, indicating the excellent internal reliability of our measurement structure. Moreover, item-total correlation can help us to check if any item in the set of tests is inconsistent with the averaged behavior of the others, and thus can be discarded. A small item-correlation value provides empirical evidence that the item is not measuring the same construct measured by the other items included. A correlation value less than 0.4 indicates that the corresponding item does not correlate very well with the scale overall and, thus, it may be dropped.

On the other hand, the convergent validity was measured by composite reliability (CR) and average variance extracted (AVE). As mentioned, we kept 23 question items with factor loading more than 0.5, presented in Table 5.1. This meets the requirements of 0.5 acceptable level, as demanded by the book [51]. Among these 23 question items, the highest factor loading is the Q20's contribution to PF, which is 0.79. Except for BI with a 0.59 composite reliability, the rest principal constructs' CA exceeds the standard level of 0.6 [58], which is a satisfying performance. As for AVE, except for "Concerns" with a 0.47 value, the six left principal constructs achieve an average variance extraction of more than 0.5 [58]. It is worth noting that the highest average variance extraction is 0.72 from SerInfQ. Based on the facts revealed above, our hypothesized model possesses both satisfactory internal reliability and robust convergent reliability. In the next section, we carry out a structural equation model for further analysis.

5.2 Structural Equation Modeling

In Figure 3.2, we presented our hypothesis on those seven principal constructs. In this section, we validated the model structure with structural equation modeling. Figure 5.1 shows the results of SEM analysis. Lavaan normally ended after 46 iterations, using Lavaan version 0.6.12. We applied maximum likelihood (ML) as the estimator, and NLMINB (Nonlinear Minimization subject to Box Constraints) as the optimization method. In total, we covered 630 responses from our *Dec2Dataset*, which results that the degree of freedom being 253. The χ^2 of our SEM model is 691.818, with the p-value of $\chi^2 < 0.001$. Moreover, CFI = 0.938 (> 0.9), TLI = 0.916 (> 0.9), RMSEA = 0.058 (< 0.08), which all surpass the recommended standards [51].

After the whole picture, we focused on the relationships inside our path structure among seven principal constructs. "Concerns" negatively influence PF (Perceived Fairness), where β = -0.139, p = 0.012. In general, a p-value less than 0.05 is statistically significant. It indicates strong evidence against the null hypothesis, as there is less than a 5% probability that the null is correct. We can claim that the correlation between "Concerns" to PF is statistically significant. Further, we have a relatively significant positive influence from SerInfQ (Service and Information Quality) to PEOU (Perceived Ease of Use) And PU (Perceived Usefulness), with respectively β = 0.981, p < 0.001 and β = 0.841, p < 0.001. These two significant factor loadings also coincide with our hypotheses, which is consistent with previous research mentioned in Chapter 2.

In the second layer, we have PEOU (Perceived Ease of Use), PF (Perceived Fairness), and PU (Perceived Usefulness), which are also the three perceived principal constructs of this study. PEOU positively influences PF, with $\beta = 0.361$, $p < 0.001$. Further, PF has a positive influence on PU, with $\beta = 0.104$, $p = 0.026$.

	Concerns	SerInfQ	PF	PEOU	PU	Att	BI
Concerns	0.687						
SerInfQ	0.000	0.848					
PF	-0.139	0.354	0.711				
PEOU	0.000	0.981	0.361	0.814			
PU	-0.014	0.878	0.401	0.862	0.827		
Att	-0.035	0.089	0.253	0.091	0.101	0.718	
BI	-0.024	0.796	0.437	0.801	0.788	0.611	0.824

Table 5.2: Inter-construct correlation matrix.

From the third and fourth layers, PF shows a significant positive effect on Att (Attitude Towards Use), with $\beta = 0.253$, $p < 0.001$. Lastly, we have three variables leading to BI (Behavioral Intentions). PEOU, Att, and PU contribute positively to BI, with $\beta = 0.466$, $p = 0.010$; $\beta = 0.535$, $p < 0.001$; and $\beta = 0.332$, respectively, $p < 0.051$. Moreover, we evaluated the discriminant validity of the principal constructs by analyzing the inter-construct correlation matrix, which is presented in Table 5.2.

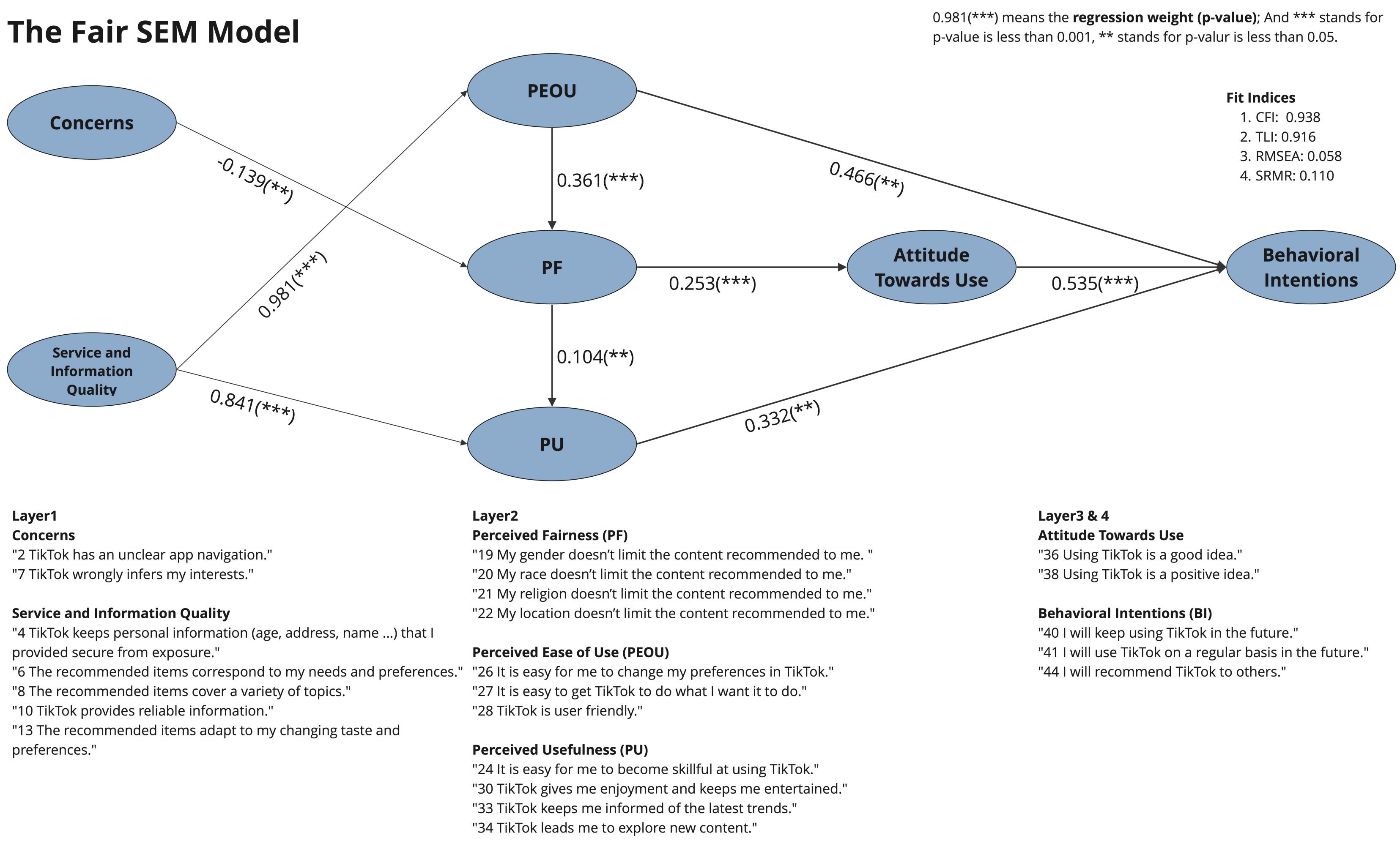

Figure 5.1: The FAIR SEM Model.

Chapter 6

Conclusion & Discussion

6.1 Conclusions

In this thesis, we discussed user perception concerning fairness issues. We investigated the position of fairness based on user perception in two-sided recommender systems. TikTok, as a perfect case study of a two-sided recommender system, gave us precious user perception from user perspectives. We thoughtfully compared and evaluated different concepts related to fairness issues. Establishing a questionnaire survey on EPFL's server, we distributed 1300 responses with Amazon MTurk. After careful selection, we obtained the *Dec2Dataset* with 630 responses. Based on this dataset, we applied structural equation modeling to investigate the relationship among perceived fairness, perceived usefulness, and perceived ease of use. During this process, we also utilized EFA and CFA to validate and justify our clusters of question items. The presented results in Chapter 5 proved the hypotheses from Figure 3.2.

According to the SEM results, the model structure met our expectations. Here, we'd like to present the SEM model. On each link, the first number is regression weight; two stars mean the p-value is less than 0.5; And three stars represent the p-value is less than 0.001. Below the figure is the question items composed of each key construct.

First, it's easy to notice the only negative link from Concerns to PF. Concerns describe users' worries and dissatisfactions. So it's reasonable that the more concerns users have, the less fair they think the recommender system is. And we found a consistent outcome that Service and Information Quality have a largely positive influence on both PEOU and PU. In layer 2, we found a positive influence from PEOU to PF. This can be because a recommender system perceived as easy to use appears more transparent to the users. You can see from the question items within PEOU, like question 26 and question 27. The more transparent a recommender system is, the more fair users think the platform is, which increases Perceived Fairness. Then PF also has a positive influence on PU. Before talking about the reasons, In 2021, we did a qualitative study focusing on the fairness issue in TikTok as well. We conducted 30 interviews to understand users' understanding of Fairness. A conclusion from our qualitative study is that PF will not directly increase users' perceived usefulness. And better PF can make users satisfied with recommended content. Thus, they will have a positive attitude towards the platform, and this can transfer into positive behavior intentions later. Based on this, it's explainable that we found a small positive influence from PF to PU, and PF positively links Attitude and links BI.

Looking deeper into Perceived Fairness. We applied EFA to compute factor loadings, which are the numerical coefficients corresponding to the directional paths connecting common factors to observed variables. Different from other constructs, we found that six question items showed a strong relationship to Perceived Fairness, with a factor loading of more than 0.6. In descending order, they respectively are race(0.742), gender(0.693), religion(0.669), social connections(0.664), age(0.660), and location(0.618). In the above SEM model, we only used 4 of them because the scree plot indicated the number of factors to remain. In multivariate statistics, a scree plot is a line plot of the eigenvalues of factors or principal components in an analysis. According to the scree plot, factors or components on the left of the "elbow" point, where the eigenvalues seem to level off is found, should be retained as significant. This result is consistent with our qualitative study. When talking about possible factors leading to Envy in RecSys, race and gender are the top two factors that are mentioned repetitively by the interviewees. And in our study, these two also significant features of Perceived Fairness from users' perceptions.

In summary, we covered seven key constructs, including "Concerns", "Service and Information Quality", "Perceived Ease of Use", "Perceived Fairness", "Perceived Usefulness", "Attitude Towards Use", and "Behavioral Intentions". We embraced the common-used constructs, such as "Perceived Ease Of Use" and Perceived Usefulness. Meanwhile, we introduced Perceived Fairness to explore the relationships among these three key constructs. We validated each key construct's statistics, Cronbach's alpha, and item-total correlation. These indices supported the statistical reliability and validity of our model analysis. We applied the structural equation modeling method to prove the consistency of our hypotheses. According to the SEM structure in Figure 5.1, The main conclusions from our research are:

- (1) Perceived Ease Of Use has a positive influence(regression weight is 0.361 with a p-value less than 0.001) on Perceived Fairness;
- (2) Perceived Fairness positively influences Perceived Usefulness;
- (3) PU and PEOU can directly increase BI, and PF can indirectly increase Bi through Attitude. These three constructs have direct and indirect positive effects on Behavioral Intentions through direct or indirect correlation.

6.2 Limitations

Further, there are several things that could be improved from our current study using the responses from MTurk.

(1) The quality of responses from MTurk is not satisfying. The rejection rate is high. After data filtering, we only got 630 out of 1300 responses. This decreases the reliability of our study.

(2) Several relationship links didn't emerge as we expected. For example, we didn't find a significant construct of the "System Quality" latent variable. Instead, we extracted the construct "Concerns" from the questions set initially for "System Quality" and "Service and Information Quality".

(3) Due to unsatisfying data quality, we exclude half of the question items to meet the convergent validity. In order to increase model reliability, we excluded many question items that might be related to the key construct. The detailed criteria are Cronbach's Alpha of each key construct should be more than 0.7, and the factor loading of each item should be more than 0.5. Specifically speaking, we have 45 main questions in our original survey questionnaire, and we got 23 questions entering our SEM model.

(4) Last, we didn't find a direct influence from PEOU to PU. However, the link from PEOU to PU is critical in the TAM model.

Hopefully, our study could bring a new perspective to researchers about the fairness problems in two-sided recommender systems. And fellow researchers could conduct further studies to add solid details to our research. We believe more attention to fairness is beneficial and necessary in recommender systems.

Bibliography

[1] Jakob Mökander and Luciano Floridi. "Ethics-Based Auditing to Develop Trustworthy AI". In: *Minds and Machines* 31.2 (June 2021), pp. 323–327. ISSN: 1572-8641. DOI: 10.1007/s11023-021-09557-8.

[2] Yongfeng Zhang, Qi Zhao, Yi Zhang, Daniel Friedman, Min Zhang, Yiqun Liu, and Shaoping Ma. "Economic Recommendation with Surplus Maximization". In: *Proceedings of the 25th International Conference on World Wide Web.* 2016, pp. 73–83.

[3] Himan Abdollahpouri and Robin Burke. "Multi-Stakeholder Recommendation and Its Connection to Multi-sided Fairness". In: *arXiv:1907.13158 [cs]* (July 2019). arXiv: 1907.13158 [cs].

[4] Gourab K Patro, Arpita Biswas, Niloy Ganguly, Krishna P. Gummadi, and Abhijnan Chakraborty. "FairRec: Two-Sided Fairness for Personalized Recommendations in Two-Sided Platforms". In: *Proceedings of The Web Conference 2020.* WWW '20. New York, NY, USA: Association for Computing Machinery, Apr. 2020, pp. 1194–1204. ISBN: 978-1-4503-7023-3. DOI: 10.1145/3366423.3380196.

[5] Tom Sühr, Asia J. Biega, Meike Zehlike, Krishna P. Gummadi, and Abhijnan Chakraborty. "Two-Sided Fairness for Repeated Matchings in Two-Sided Markets: A Case Study of a Ride-Hailing Platform". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.* KDD '19. New York, NY, USA: Association for Computing Machinery, July 2019, pp. 3082–3092. ISBN: 978-1-4503-6201-6. DOI: 10.1145/3292500.3330793.

[6] Lequn Wang and Thorsten Joachims. "User Fairness, Item Fairness, and Diversity for Rankings in Two-Sided Markets". In: *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval.* 2021, pp. 23–41.

[7] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. "Fairness-Aware Ranking in Search & Recommendation Systems with Application to Linkedin Talent Search". In: *Proceedings of the 25th Acm Sigkdd International Conference on Knowledge Discovery & Data Mining.* 2019, pp. 2221–2231.

[8] Nour Nasr. *TikTok: Key Figures and Statistics in the UK and in the World in 2022*. https://blog.digimind.com/en/agencies/tiktok-key-figures-and-statistics-in-the-uk-and-in-the-world-in-2022.

[9] United Nations. *World Population Prospects 2022: Summary of Results*. United Nations, Aug. 2022. ISBN: 978-92-1-001438-0. DOI: 10.18356/9789210014380.

[10] Kemp Simon. *Digital 2022: April Global Statshot Report [Page 177: Profile of TikTok's Global Advertising Audience by Age and Gender (TikTok Demographics)]*. https://datareportal.com/reports/digital-2022-april-global-statshot. Apr. 2022.

[11] Robert K. Merton. "The Matthew Effect in Science: The Reward and Communication Systems of Science Are Considered." In: *Science* 159.3810 (1968), pp. 56–63.

[12] Julia Dressel and Hany Farid. "The Accuracy, Fairness, and Limits of Predicting Recidivism". In: *Science Advances* (Jan. 2018). DOI: 10.1126/sciadv.aao5580.

[13] Alexandra Chouldechova. "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments". In: *Big Data* 5.2 (June 2017), pp. 153–163. ISSN: 2167-6461. DOI: 10.1089/big.2016.0047.

[14] Zuohui Fu, Yikun Xian, Ruoyuan Gao, Jieyu Zhao, Qiaoying Huang, Yingqiang Ge, Shuyuan Xu, Shijie Geng, Chirag Shah, Yongfeng Zhang, and Gerard de Melo. "Fairness-Aware Explainable Recommendation over Knowledge Graphs". In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: Association for Computing Machinery, July 2020, pp. 69–78. ISBN: 978-1-4503-8016-4.

[15] Robin Burke. "Multisided Fairness for Recommendation". In: *arXiv:1707.00093 [cs]* (July 2017). arXiv: 1707.00093 [cs].

[16] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. "Fairness Through Awareness". In: *arXiv:1104.3913 [cs]* (Nov. 2011). arXiv: 1104.3913 [cs].

[17] Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. "Equity of Attention: Amortizing Individual Fairness in Rankings". In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. SIGIR '18. New York, NY, USA: Association for Computing Machinery, June 2018, pp. 405–414. ISBN: 978-1-4503-5657-2. DOI: 10.1145/3209978.3210063.

[18] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. "The Unfairness of Popularity Bias in Recommendation". In: *arXiv:1907.13286 [cs]* (Sept. 2019). arXiv: 1907.13286 [cs].

[19] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. "Controlling Popularity Bias in Learning-to-Rank Recommendation". In: *Proceedings of the Eleventh ACM Conference on Recommender Systems*. Como Italy: ACM, Aug. 2017, pp. 42–46. ISBN: 978-1-4503-4652-8. DOI: 10.1145/3109859.3109912.

[20] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. "Managing Popularity Bias in Recommender Systems with Personalized Re-Ranking". In: *The Thirty-Second International Flairs Conference.* May 2019.

[21] Gediminas Adomavicius and YoungOk Kwon. "Improving Aggregate Recommendation Diversity Using Ranking-Based Techniques". In: *IEEE Transactions on Knowledge and Data Engineering* 24.5 (May 2012), pp. 896–911. ISSN: 1558-2191. DOI: 10.1109/TKDE.2011.15.

[22] Haolun Wu, Chen Ma, Bhaskar Mitra, Fernando Diaz, and Xue Liu. "Multi-FR: A Multi-Objective Optimization Method for Achieving Two-sided Fairness in E-commerce Recommendation". In: *arXiv:2105.02951 [cs]* (May 2021). arXiv: 2105.02951 [cs].

[23] Yao Wu, Jian Cao, Guandong Xu, and Yudong Tan. "TFROM: A Two-sided Fairness-Aware Recommendation Model for Both Customers and Providers". In: *arXiv:2104.09024 [cs]* (Apr. 2021). arXiv: 2104.09024 [cs].

[24] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. "Inherent Trade-Offs in the Fair Determination of Risk Scores". In: *arXiv:1609.05807 [cs, stat]* (Nov. 2016). arXiv: 1609.05807 [cs, stat].

[25] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P. Gummadi, and Adrian Weller. "Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 32.1 (Apr. 2018). ISSN: 2374-3468.

[26] Duncan Karl Foley. *Resource Allocation and the Public Sector.* Yale University, 1966.

[27] Maria-Florina Balcan, Travis Dick, Ritesh Noothigattu, and Ariel D. Procaccia. "Envy-Free Classification". In: *arXiv:1809.08700 [cs, stat]* (Sept. 2020). arXiv: 1809.08700 [cs, stat].

[28] Berk Ustun, Yang Liu, and David Parkes. "Fairness without Harm: Decoupled Classifiers with Preference Guarantees". In: *Proceedings of the 36th International Conference on Machine Learning.* PMLR, May 2019, pp. 6373–6382.

[29] Tony Ahn, Seewon Ryu, and Ingoo Han. "The Impact of Web Quality and Playfulness on User Acceptance of Online Retailing". In: *Information & Management* 44.3 (Apr. 2007), pp. 263–275. ISSN: 0378-7206. DOI: 10.1016/j.im.2006.12.008.

[30] Kun Zhao, Eamonn Ferguson, and Luke D. Smillie. "When Fair Is Not Equal: Compassion and Politeness Predict Allocations of Wealth Under Different Norms of Equity and Need". In: *Social Psychological and Personality Science* 8.8 (Nov. 2017), pp. 847–857. ISSN: 1948-5506. DOI: 10.1177/1948550616683018.

[31] Paul G. Curran. “Methods for the Detection of Carelessly Invalid Responses in Survey Data”. In: *Journal of Experimental Social Psychology.* Rigorous and Replicable Methods in Social Psychology 66 (Sept. 2016), pp. 4–19. ISSN: 0022-1031. DOI: 10.1016/j.jesp.2015.07.006.

[32] Fred D. Davis. “Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology”. In: *MIS Quarterly* 13.3 (1989), pp. 319–340. ISSN: 0276-7783. DOI: 10.2307/249008.

[33] Martin Fishbein and Icek Ajzen. “Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research”. In: *Philosophy and Rhetoric* 10.2 (1977).

[34] Ji-Won Moon and Young-Gul Kim. “Extending the TAM for a World-Wide-Web Context”. In: *Information & Management* 38.4 (Feb. 2001), pp. 217–230. ISSN: 0378-7206. DOI: 10.1016/S0378-7206(00)00061-6.

[35] Tien T. Nguyen, Pik-Mai Hui, F. Maxwell Harper, Loren Terveen, and Joseph A. Konstan. “Exploring the Filter Bubble: The Effect of Using Recommender Systems on Content Diversity”. In: *Proceedings of the 23rd International Conference on World Wide Web.* WWW ’14. New York, NY, USA: Association for Computing Machinery, Apr. 2014, pp. 677–686. ISBN: 978-1-4503-2744-2. DOI: 10.1145/2566486.2568012.

[36] Virginie Do, Sam Corbett-Davies, Jamal Atif, and Nicolas Usunier. *Online Certification of Preference-Based Fairness for Personalized Recommender Systems.* Jan. 2022. arXiv: 2104.14527 [cs, stat].

[37] Michael P. Kim, Aleksandra Korolova, Guy N. Rothblum, and Gal Yona. “Preference-Informed Fairness”. In: *arXiv preprint arXiv:1904.01793* (2019). arXiv: 1904.01793.

[38] Muhammad Bilal Zafar, Isabel Valera, Manuel Rodriguez, Krishna Gummadi, and Adrian Weller. “From Parity to Preference-Based Notions of Fairness in Classification”. In: *Advances in neural information processing systems* 30 (2017).

[39] James L. Gibson. *Organizations: Behavior, Structure, Processes.* McGraw-Hill, 2012. ISBN: 978-0-07-108641-7.

[40] Harry Charalambos Triandis. *Attitude and Attitude Change.* Wiley Foundations in Social Psychology. New York: John Wiley & Sons, Inc, 1971. ISBN: 978-0-471-88831-4 978-0-471-88830-7.

[41] M. A. AlKhaldi and I. M. AlJabri. “The Relationship of Attitudes to Computer Utilization: New Evidence from a Developing Nation”. In: *Computers in Human Behavior* 14.1 (Jan. 1998), pp. 23–42. ISSN: 0747-5632. DOI: 10.1016/S0747-5632(97)00030-7.

[42] Viswanath Venkatesh, Michael G. Morris, Gordon B. Davis, and Fred D. Davis. “User Acceptance of Information Technology: Toward a Unified View”. In: *MIS Quarterly* 27.3 (2003), pp. 425–478. ISSN: 0276-7783. DOI: 10.2307/30036540.

[43] Pearl Pu, Li Chen, and Rong Hu. "A User-Centric Evaluation Framework for Recommender Systems". In: *Proceedings of the Fifth ACM Conference on Recommender Systems*. RecSys '11. New York, NY, USA: Association for Computing Machinery, Oct. 2011, pp. 157–164. ISBN: 978-1-4503-0683-6. DOI: 10.1145/2043932.2043962.

[44] Brahim Zarouali, Evert Van den Broeck, Michel Walrave, and Karolien Poels. "Predicting Consumer Responses to a Chatbot on Facebook". In: *Cyberpsychology, Behavior, and Social Networking* 21.8 (Aug. 2018), pp. 491–497. ISSN: 2152-2715. DOI: 10.1089/cyber.2017.0518.

[45] *Amazon Mechanical Turk*. https://www.mturk.com/.

[46] Herman Aguinis, Isabel Villamor, and Ravi S. Ramani. "MTurk Research: Review and Recommendations". In: *Journal of Management* 47.4 (Apr. 2021), pp. 823–837. ISSN: 0149-2063. DOI: 10.1177/0149206320969787.

[47] Sheryl L. Walter, Scott E. Seibert, Daniel Goering, and Ernest H. O'Boyle. "A Tale of Two Sample Sources: Do Results from Online Panel Data and Conventional Data Converge?" In: *Journal of Business and Psychology* 34.4 (Aug. 2019), pp. 425–452. ISSN: 0889-3268. DOI: 10.1007/s10869-018-9552-y.

[48] Christopher O. L. H. Porter, Ryan Outlaw, Jake P. Gale, and Thomas S. Cho. "The Use of Online Panel Data in Management Research: A Review and Recommendations". In: *Journal of Management* 45.1 (Jan. 2019), pp. 319–344. ISSN: 0149-2063. DOI: 10.1177/0149206318811569.

[49] Richard N. Landers and Tara S. Behrend. "An Inconvenient Truth: Arbitrary Distinctions Between Organizational, Mechanical Turk, and Other Convenience Samples". In: *Industrial and Organizational Psychology* 8.2 (June 2015), pp. 142–164. ISSN: 1754-9426, 1754-9434. DOI: 10.1017/iop.2015.13.

[50] Theodoros A. Kyriazos. "Applied Psychometrics: Sample Size and Sample Power Considerations in Factor Analysis (EFA, CFA) and SEM in General". In: *Psychology* 9.8 (Aug. 2018), pp. 2207–2230. DOI: 10.4236/psych.2018.98126.

[51] Joseph F. Hair | Barry J. Babin | Rolph E. Anderson | William C. Black. *Multivariate Data Analysis, 8th Edition*. CENGAGE INDIA, Jan. 2018. ISBN: 978-93-5350-135-8.

[52] Darren George and Paul Mallery. *SPSS for Windows Step by Step: A Simple Guide and Reference 18.0 Update*. 11th edition. Boston, MA: Taylor & Francis, Jan. 2010. ISBN: 978-0-205-01124-7.

[53] Robert A. Peterson. "A Meta-analysis of Cronbach's Coefficient Alpha". In: *Journal of Consumer Research* 21.2 (Sept. 1994), pp. 381–391. ISSN: 0093-5301. DOI: 10.1086/209405.

[54] Eduard Cristobal, Carlos Flavián, and Miguel Guinalíu. "Perceived E-service Quality (PeSQ): Measurement Validation and Effects on Consumer Satisfaction and Web Site Loyalty". In: *Managing Service Quality: An International Journal* 17.3 (Jan. 2007), pp. 317–340. ISSN: 0960-4529. DOI: 10.1108/09604520710744326.

[55] Eleanor T. Loiacono, Richard T. Watson, and Dale L. Goodhue. "WebQual: A Measure of Website Quality". In: *Marketing theory and applications* 13.3 (2002), pp. 432–438.

[56] Julie E. Francis and Lesley White. "PIRQUAL: A Scale for Measuring Customer Expectations and Perceptions of Quality in Internet Retailing". In: *K. Evans & L. Scheer (Eds.)* (2002), pp. 263–270.

[57] Soyoung Kim and Leslie Stoel. "Apparel Retailers: Website Quality Dimensions and Satisfaction". In: *Journal of retailing and consumer services* 11.2 (2004), pp. 109–117.

[58] Claes Fornell and David F. Larcker. "Evaluating Structural Equation Models with Unobservable Variables and Measurement Error". In: *Journal of marketing research* 18.1 (1981), pp. 39–50.

Appendix A

Optional Readings

A.1 The FAIR model survey

We are scientists from EPFL, a leading engineering university in Europe. We are conducting a large scale user survey to understand your interaction experiences with TikTok. We are especially interested in knowing if this recommendation platform seems fair to you. We are going to ask you to take a moment to reflect on your recent user experience with TikTok, the recommendations you have received lately, their quality, the service quality provided, your privacy concerns, and whether you feel in control of the content you receive. You will then answer 45 questions and this survey should take about 30 minutes.

If you participate in this project, you will receive $12 for your participation. However, if your experience with TikTok is less than 6 months, or you are not a fluent speaker of English, you will not be qualified.

All data collected in this project is subject to strict data protection rules. If you decide to join, you will find a more detailed information sheet about why this research is being carried out and what it means to you.

Demographic questions:

Pre-survey questions [6 questions]

1. Please specify your gender:

- female
- male
- other

2. Please specify your age group:

- 18-24
- 25-34
- 35-44
- 45-54
- 55-65

3. Please specify your English fluency:

- native
- fluent
- other

4. Please specify your profession:

- student
- office worker
- independent
- retired
- other

5. Please specify your TikTok experience:

- over 6 months of usage
- 0-6 months of usage
- other

6. Please specify your weekly usage of TikTok:

- longer than 2 hours
- about 1-2 hours
- less than 1 hour
- less than 30 mins

Main questions:

System quality [6 questions]

1. TikTok has an appropriate style of design for site type.
2. It's easy to navigate on TikTok to find what I want.
3. TikTok has unclear app navigation.
4. TikTok has a fast response time.
5. TikTok keeps personal information (age, address, name ...) that I provided secure from

exposure.

6. TikTok creates an enjoyable audiovisual experience.

Service (and Information) quality [10 questions]

7. The recommended items correspond to my needs and preferences.
8. TikTok wrongly infers my interests.
9. The recommended items cover a variety of topics.
10. TikTok recommended items are repetitive.
11. TikTok provides reliable information.
12. The recommended items are novel.
13. I can easily change the way TikTok recommends content to me.
14. The recommended items adapt to my changing taste and preferences.
15. TikTok recommends me content based on inferred information, such as gender, location, profession, etc.
16. TikTok displays too many advertisements.

Perceived Fairness [8 questions]

17. The content from TikTok controls my preferences and taste.
18. The content recommended to me is populated by trending videos.
19. My age doesn't limit the content recommended to me.
20. My gender doesn't limit the content recommended to me.
21. My race doesn't limit the content recommended to me.
22. My religion doesn't limit the content recommended to me.
23. My location doesn't limit the content recommended to me.
24. My social connections don't limit the content recommended to me.

Perceived ease of use [5 questions]

25. It is easy for me to become skillful at using TikTok.
26. Using TikTok requires a lot of mental effort.
27. It is easy for me to change my preferences in TikTok.
28. It is easy to get TikTok to do what I want it to do.
29. TikTok is user-friendly.

Perceived usefulness [7 questions]

30. When using TikTok, I do not realize the time elapsed.
31. TikTok gives me enjoyment and keeps me entertained.
32. Using TikTok is dull and boring.
33. TikTok stimulates my curiosity.

34. TikTok keeps me informed of the latest trends.
35. TikTok leads me to explore new content.
36. TikTok updates me about my friends' lives.

Attitude toward use [4 questions]

37. Using TikTok is a good idea.
38. Using TikTok is a bad idea.
39. Using TikTok is a positive idea.
40. Using TikTok is a wise idea.

Behavioral intention to use (Users' adoption) [5 questions]

41. I will keep using TikTok in the future.
42. I will use TikTok on a regular basis in the future.
43. I don't intend to use TikTok in the future.
44. I will use this site rather than other platforms, for example, YouTube, Instagram, or Facebook.
45. I will recommend TikTok to others.

A.2 Response Distributions

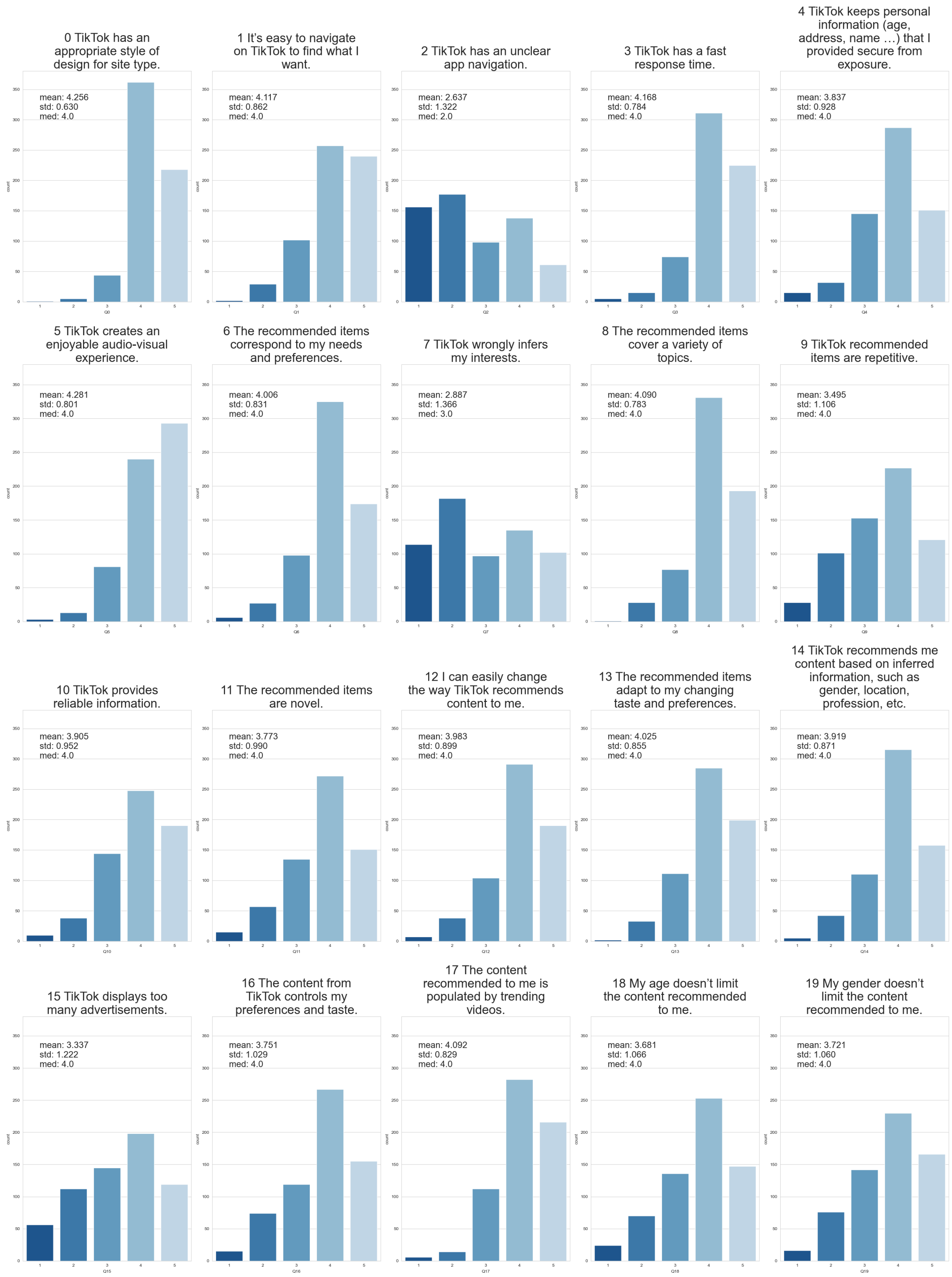

Figure A.1: The distributions of questions [0-19], with statistics.

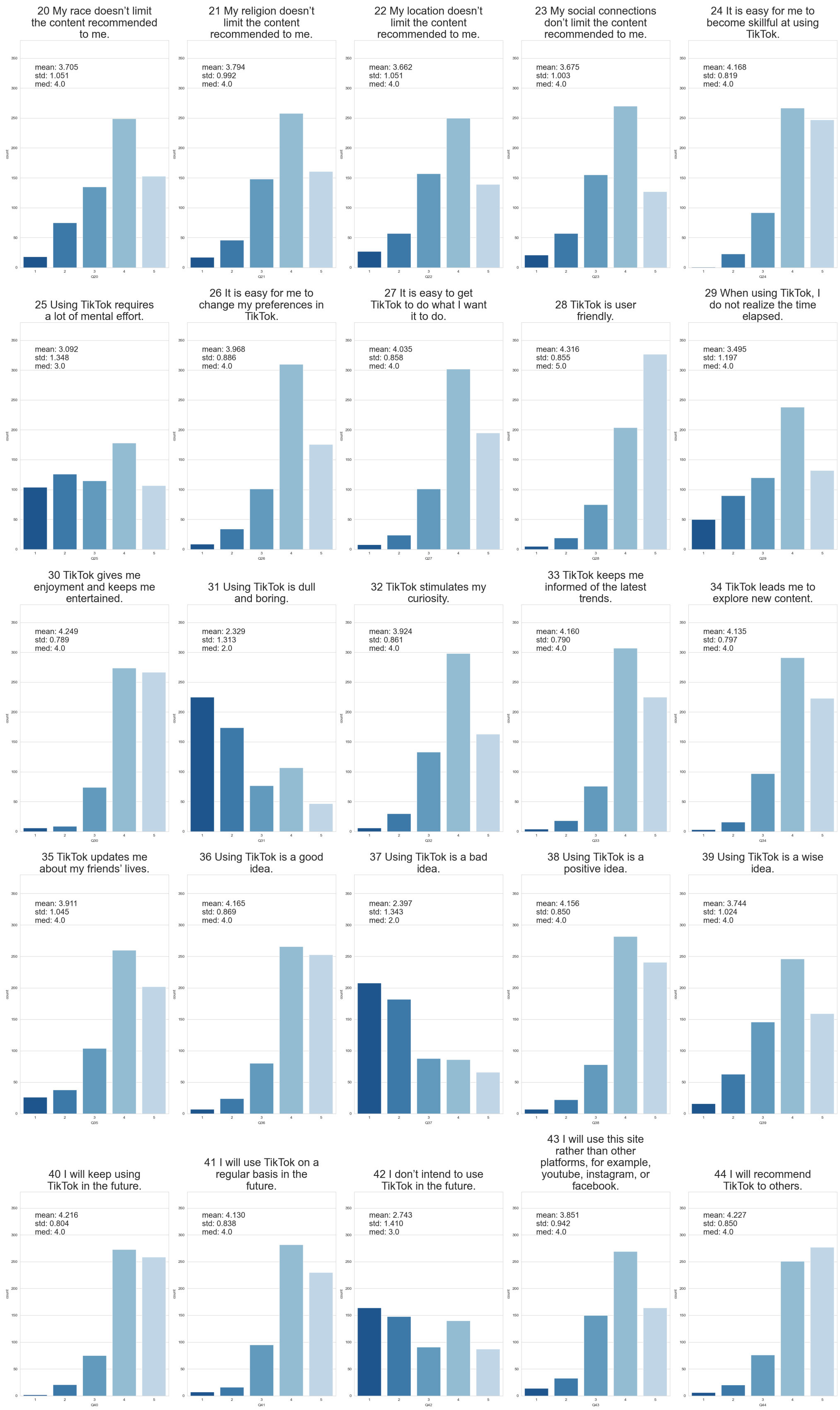

Figure A.2: The distributions of questions [20-45], with statistics.

A.3 Response Correlation Map

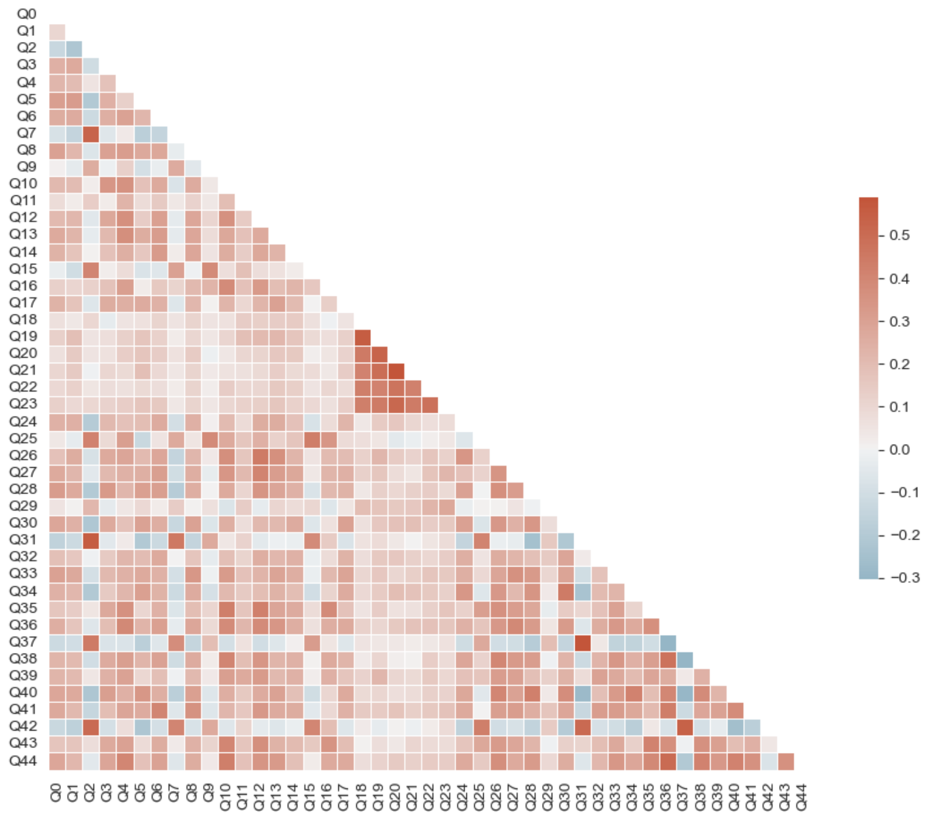

Figure A.3: The correlation map of questions [0-44].

Construct	N	Statistics			Internal reliability		Convergent validity		
		Mean	SD	Median	Cronbach's alpha (0.6)	Item-total correlation (0.5)	Factor loading (0.5)	Composite reliability (0.6)	Average variance extracted (0.5)
System Quality	4				0.75			0.76	0.54
Q0		4.17	0.62	4.00		0.67	0.53		
Q1		4.15	0.88	4.00		0.81	0.76		
Q2		2.03	0.94	2.00		0.85	0.83		
Q5		4.35	0.75	4.00		0.69	0.53		
Service & Information Quality	6				0.77			0.78	0.63
Q6		3.93	0.81	4.00		0.73	0.69		
Q7		2.28	1.05	2.00		0.68	0.60		
Q8		3.90	0.85	4.00		0.61	0.54		
Q9		2.85	1.03	3.00		0.72	0.58		
Q12		3.48	1.08	4.00		0.68	0.57		
Q13		3.83	0.88	4.00		0.69	0.66		
Perceived fairness	6				0.87			0.87	0.47
Q18		3.72	1.02	4.00		0.72	0.62		
Q19		3.60	1.05	4.00		0.81	0.77		
Q20		3.71	1.04	4.00		0.83	0.81		
Q21		3.76	0.98	4.00		0.77	0.72		
Q22		3.59	1.07	4.00		0.75	0.68		
Q23		3.65	0.93	4.00		0.79	0.75		
Perceived ease of use	4				0.75			0.71	0.62
Q24		4.00	0.94	4.00		0.73	0.58		
Q26		3.66	0.98	4.00		0.79	0.66		
Q27		3.72	0.98	4.00		0.79	0.61		
Q28		4.20	0.83	4.00		0.71	0.60		
Perceived usefulness	5				0.80			0.78	0.58
Q30		4.28	0.74	4.00		0.81	0.79		
Q31		1.75	0.83	2.00		0.76	0.76		
Q32		3.92	0.81	4.00		0.71	0.53		
Q33		4.09	0.78	4.00		0.73	0.54		
Q34		4.16	0.73	4.00		0.73	0.57		
Attitude toward use	4				0.88			0.88	0.34
Q36		3.76	1.03	4.00		0.91	0.91		
Q37		2.14	1.02	2.00		0.88	0.86		
Q38		3.82	0.92	4.00		0.86	0.82		
Q39		3.38	1.02	3.00		0.77	0.63		
Behavioral intentions	4				0.83			0.82	0.47
Q40		4.26	0.78	4.00		0.88	0.83		
Q41		4.08	0.85	4.00		0.84	0.75		
Q42		1.92	1.00	2.00		0.81	0.66		
Q44		3.95	0.97	4.00		0.77	0.66		

Table A.1: Summary of internal reliability and convergent validity.