INTELLIGENT
TECHNOLOGIES
IN EDUCATION

Choice Diversity in Educational Recommender Systems: User Perspectives

Robert Wesley Songer[a], Tomohito Yamamoto[a]

[a] Department of Information and Computer Engineering, Kanazawa Institute of Technology, Kanazawa, Japan

Abstract

Educational recommender systems (ERSs) traditionally prioritize prediction accuracy, often overlooking the impact of recommendation diversity on user satisfaction. This study aims to understand how recommendation diversity and user psychological traits, such as need for cognitive closure, affect user satisfaction and preferences in ERSs. In two experiments involving university students, we analyzed subjective perceptions of recommendation qualities—including accuracy, novelty, and usefulness—and evaluated the effects of psychological priming on user evaluations. The results reveal that user satisfaction depends not only on perceived accuracy but also on the interplay of diversity, perceived usefulness, and novelty. Furthermore, priming users to consider accuracy or diversity prior to using the system appeared to mitigate the influence of psychological traits, resulting in more consistent evaluations. These findings highlight the potential of task-based strategies, psychological trait personalization, and priming for designing ERSs that foster more effective and satisfying learning experiences.

Editors

Section: Educational Technology

Co-Editor-in-Chief: Dr Marios Kremantzis
Associate Editor: Dr Xue Zhou

Publication

Received: 24 August 2024
Revised: 10 November 2024
Accepted: 16 December 2024
Online: 17 December 2024

Copyright

© by the authors, in its year of first publication. This publication is an open access publication under the Creative Commons Attribution CC BY-ND 4.0 license.

Keywords

Educational Technology; Recommender Systems; Recommendation Diversity; Decision Making; Cognitive Closure

Citation

Songer, RW., & Yamamoto, T. (2024). Choice Diversity in Educational Recommender Systems: User Perspectives, *Intelligent Technologies in Education,* Advanced Online Publication

Introduction

In technology-enhanced learning, recommender systems have been increasing in popularity with advances in big data and artificial intelligence techniques. Educational recommender systems (ERSs) share a unique set of goals for efficiently and effectively supporting the learning process (Manouselis et al., 2012). The most common goal of ERSs is to help learners find learning resources such as content, activities, or sequences of items (Drachsler et al., 2015), essentially acting as a decision support system that augments the user's ability to search and choose from a large pool of available options.

Much of ERS research has conventionally focused on outcome-oriented evaluation metrics such as algorithmic prediction accuracy and user satisfaction, essentially following the same research approach as that of commercial applications (Erdt nee Anjorin et al., 2015; da Silva et al., 2022). Less common are studies that measure user perceptions of recommendation qualities such as usefulness, novelty, and diversity (Marante et al., 2022; Deschênes, 2020). The majority of ERS studies evaluate recommendations with prediction accuracy metrics (e.g., precision and recall), while those that report user-centered measures generally do not report user or recommendation qualities other than satisfaction. That is, rather than examining the learner's experiences from using the system, evaluations more commonly assess the system's ability to predict choices that will satisfy the learner without understanding why. This approach seems to lack a focus on user satisfaction, which stems not so much from the final item selected but rather from the process of exploring and selecting options—a process that can benefit from recommendations with qualities such as usefulness, novelty, diversity, and serendipity (Fazeli et al., 2018). User-centered evaluations provide valuable opportunities to understand key behaviors that recommender systems aim to support (Chen et al., 2013a). To this end, recommendations based on evaluation metrics such as item diversity and novelty have been proposed for their ability to expand user perspectives (Zhou et al., 2010). Researchers have also explored user qualities related to a tolerance or desire for unexpected recommendations, such as their level of curiosity (Zhao & Lee, 2016) or Big Five personality traits (Chen et al., 2013b; Wu & Chen, 2013), and proposed recommendation frameworks based on their findings.

Especially in education where decision making and learning occur side-by-side, recommender systems ought to be designed with inherent aspects of the learning process in mind. While recent ERS research has gained much ground in this regard, there are still gaps in the areas of application and methods of recommendation, such as the use of hybrid techniques that intelligently combine user information (da Silva et al., 2022; Urdaneta-Ponte et al., 2021). Investigating the connection between recommendation qualities and user experiences is one such avenue of research that needs to be further explored. Other issues related to general influences surrounding the presentation of recommendations, such as priming the user with some stimuli prior to rating the recommendations they receive, warrant the development of user-centric evaluation techniques (Jameson et al., 2015).

The present study addresses gaps in the literature between by investigating the following questions:

- **RQ1**: What are the effects of recommendation diversity on user satisfaction and preferences in educational recommender systems?
- **RQ2**: How do psychological traits, such as decisiveness and need for predictability, influence user evaluations of recommendation qualities?
- **RQ3**: How does priming users to focus on accuracy or diversity affect their perception of recommendations?

Answering these questions will provide valuable insights into ERS design, enabling the development of systems that promote more effective and satisfying user experiences through diverse and tailored recommendations.

To this end, we compare subjective user perceptions of recommendation qualities, such as diversity, novelty, and usefulness, to algorithmic measurements of diversity for the same sets of items. The study includes two separate experiments in which students were asked to evaluate the recommendations of an original ERS used in a task to search for new study topics. The first experiment examines their experiences when given recommendation lists with varying degrees of diversity without any prior knowledge of the recommendation algorithm. The second experiment compares user-stated preferences for recommendation diversity and their perceptions of recommendations generated without an algorithmic bias for accuracy. In both experiments, participants responded to questionnaires that were intended to measure both user qualities and user evaluations of recommendation qualities. By examining psychological dispositions and perceived qualities of the user, the present study provides evidence for a link between specific aspects of the user and their experience of recommendations.

Literature

This study examines constructs related to both user and recommendation qualities. The user qualities include preferences for recommendation diversity, decisiveness, and need for predictability. The recommendation qualities include accuracy, diversity, novelty, usefulness, and satisfaction. In recommender systems research, “accuracy” typically refers to the ability of a system to predict what the user will choose; however, we use the term to describe both the statistical measure of similarity between recommendation items and user perceptions of similarity, which we explain further in the Methods section.

Recommendation Qualities

The gold standard evaluation metric for measuring the effectiveness of recommender systems has conventionally been characterized by accuracy metrics based on the assumption that users prefer items with traits that are similar to those they have chosen before (Gunawardana & Shani, 2015). While this has led to several important advances in prediction techniques, researchers have also recognized that an emphasis on prediction accuracy saturates recommendations with similar items and makes systems too predictable for the user (Amer-Yahia et al., 2009). This problem has become known as the “filter bubble” (Pariser, 2011), especially in the context of search engines and social media sites. In response, researchers have proposed several “beyond-accuracy” metrics (Kaminskas & Bridge, 2016) for their potential to improve the diversity of recommendations, provide users with opportunities to discover new items, and generally broaden users’ horizons (Zhou et al., 2010; Matt et al., 2014).

While accuracy can be used to describe the degree of similarity between item traits, diversity represents differences between traits and is calculated by comparing all the items in a given set (Castells et al., 2015). When calculating the degree of difference for single items, the metric is more commonly referred to as novelty and can be calculated by comparing one item to other items or to aspects of the user’s profile that are assumed to represent their preferences. Novelty is commonly measured in user-centered studies that attempt to identify which qualities provide value to the user. Its relationships with other qualities have been studied with respect to recommendation qualities such as diversity, coverage, perceived usefulness, and serendipity (Kaminskas & Bridge, 2016; Pu & Chen, 2011; Kotkov et al., 2018; Chen et al., 2019), as well as user qualities of satisfaction, preference, enjoyment, and curiosity (Matt et al., 2014; Kotkov et al, 2018; Chen et al., 2019; Wieland et al., 2021; Maccatrozzo et al., 2017). Many findings suggest that benefits come from nuanced combinations of qualities rather than from any one quality in particular.

The qualities of usefulness and satisfaction are commonly measured with user-centered reports; however, ERS studies tend to measure satisfaction more often than usefulness (da Silva et al., 2022), especially for systems intended to enhance student agency (Deschênes, 2020). It has long been argued that user satisfaction does not come from accurate predictions alone (McNee et al., 2006), and this claim has been backed up by empirical findings (Fazeli et al., 2018). Some studies suggest that satisfaction is connected to serendipity in certain contexts (Chen et al., 2019; Wieland et al., 2021; Lutz et al., 2017), although serendipity as a concept is ambiguous and difficult to convey in certain languages (e.g., Said et al., 2013). Kotkov et al. (2018) reviewed the constructs of serendipity used in research and discovered that, despite having several manifestations, they generally include the qualities of relevance, novelty, and unexpectedness. In the same study, they investigated which aspects of serendipity are the most important for user satisfaction in the context of a movie recommender system. Their findings suggest that novelty has a positive effect in certain cases, while unexpectedness is more nuanced and can negatively affect satisfaction. For example, the user may be less satisfied with the performance of the recommender system when they receive a recommendation that they did not expect to be relevant to them. Pu et al. investigated satisfaction through the lens of perceived usefulness and found that it was correlated with a combination of accuracy and novelty in one study (Pu & Chen, 2011), while in another study, it was correlated with diversity only when diversity enhanced the context in which the user was making their decision (Pu et al., 2012).

Due to these highly nuanced interactions between satisfaction and other perceived recommendation qualities, the current study treats user satisfaction as an overarching outcome to be evaluated alongside user perceptions of other qualities such as usefulness, novelty, and accuracy. The relationships between these qualities and satisfaction are summarized in **Table 1**.

Table 1

User perceived qualities and their relationships to overall satisfaction

Construct	Description	Relation to Satisfaction
Accuracy	The perceived degree of relevance or similarity to expected items	Improves satisfaction by providing practical value or familiarity
Novelty	The degree to which recommendations are unfamiliar or unexpected	Increases satisfaction for users who seek exploration
Usefulness	The perceived practical value of recommendations	Positively influences satisfaction, especially in practical contexts

User Qualities

User preferences have long been used as an approach in the design of recommender systems, but doing so requires accurately modeling a user's desirable traits (Gunawardana & Shani, 2015). This technique may define the content preferences of users in terms of domain-specific details, such as movie genres, effectively limiting its broader applicability. Studies that have looked more broadly at user preferences examined the relative qualities of recommendations themselves rather than content of item choices. Such preferences can generally be classified as either attraction to items perceived as having desirable qualities or aversion to items perceived as lacking such qualities (Lu et al, 2014) and may be measured from behavioral data, such as user engagement and effort levels (Mehrotra et al., 2020). For example, preference for novelty may include a desire for newness as well as a tolerance or receptivity for unexpectedness. Generalized preferences such as these have been further evidenced by a link between user satisfaction and recommendation qualities, albeit subject to the decision-making context (Chen et al., 2013; Matt et al., 2014; Lutz et al., 2017). In addition, preferences for novelty have been shown to vary between different users and even change for the same user over time (Kapoor et al., 2015; Alhijawi et al., 2022).

Individual qualities that influence a person's openness to receiving new information are described by the psychological theory of need for cognitive closure (NCC). NCC has been defined as the "individual's desire for a firm answer to a question and an aversion toward ambiguity" (Kruglanski & Webster, 1996, p. 264). It describes the processes of "seizing"—

hastily accepting a choice or belief when no prior one is available—and “freezing”—refusing to accept new information when it conflicts with one’s current beliefs. These concepts are reflected in different types of curiosity (Litman, 2010) as well as individual susceptibility to persuasion under various circumstances (Kruglanski & Webster, 1993). The theory and its associated need for closure scale (NCS) were created by integrating several findings from social psychology research and combining them into a scale for measurement. The scale aggregates five subscales, namely, need for order, desire for predictability, decisiveness, discomfort with ambiguity, and closemindedness (Webster & Kruglanski, 1994). Several studies have used the scale to measure NCC and explore its relation to other psychological traits, such as information processing behaviors. Notably, when seeking information, individuals with high NCC have been found to perform more selective processing than their low NCC counterparts when under pressure from imposed time limits or high information loads (Choi et al., 2008; Kardes et al., 2004).

The NCS subscales of decisiveness and need for predictability are particularly interesting with regard to recommendations, as they relate to psychological traits such as attitudes toward ambiguity, tolerance for uncertainty, independence, and individualism. High decisiveness has been linked to a need for complexity/novelty, while a high need for predictability has been linked to discomfort with ambiguity (Hitsuwari & Nomura; 2021). In the face of uncertainty, individuals with high decisiveness may be less likely to hesitate, while those with a high need for predictability may be more likely to experience stress (Barenbaum et al., 2008). Decisiveness has also been linked to independence, individualism, and functional impulsivity, while the need for predictability has been linked to interdependence, collectivism, and low dysfunctional impulsivity (Suzuki & Sakurai, 2001). This apparent contradiction between these two subscales gave rise to controversy surrounding the original items for decisiveness in the NCS. In the original wording, the items expressed not just the need for but also the ability to make quick, unambiguous decisions. Neuberg et al. (Neuberg et al., 1997) further noted that the decisiveness subscale correlates negatively with all other NCS subscales and suggested that the NCS is better suited as a multidimensional tool rather than a linear score aggregator. This perspective was debated but later validated to an extent when a study by Roets and Van Hiel (2007) provided empirical evidence confirming the poor relatedness of decisiveness with the other NCS subscales. In light of these findings, the present study examines decisiveness and need for predictability scores separately but also provides cumulative scores for comparison.

Method

Both experiments asked student participants to use and evaluate an experimental system, called the Topics Recommender, in a task with the aim of producing a list of desirable study topics. The participants in both experiments were Japanese graduate and undergraduate students recruited based on their voluntary availability within the Department of Information and Computer Engineering at a Japanese university. None of the participants had any prior knowledge of or experience with the Topics Recommender system, ensuring a neutral starting point for evaluating the recommendations. All participants were given uniform instructions in an identical controlled laboratory environment for each experiment, including time constraints and assistance from a researcher when necessary. All participants completed the tasks within a single session and in the same environment in order to minimize external influences. Although participants may have varied in their learning preferences or exposure to study topics, this variability is considered a natural component of the user experience. No specific controls for learning preferences or familiarity with study topics were implemented as the goal was to capture the range of responses representative of typical student interactions with an ERS. This approach reflects the study's emphasis on capturing authentic user experiences, where such variability reflects the diverse backgrounds of real-world ERS users. By focusing on typical interactions, the study provides insights for designing systems that adapt to a wide range of user needs and contexts.

In Experiment 1, participants first completed the search task before answering a questionnaire containing items related to perceived recommendation qualities and general preferences for recommendation diversity. After two weeks, the participants were contacted again and asked to complete the need for cognitive closure questionnaire. In Experiment 2, participants first answered a pretest NCC questionnaire before using the Topics Recommender system. After they completed the search task, they answered a posttest questionnaire about perceived recommendation qualities. In order to examine differences in reported preferences before and after using the system, half of the participants received the item regarding preferences for recommendation diversity in the pretest questionnaire, while the other half received the item in the posttest questionnaire. All the questionnaires were given in Japanese, and responses were linked to each participant’s session data in the Topics Recommender system for later analysis. Further details about the system, task, and flow are in the following sections.

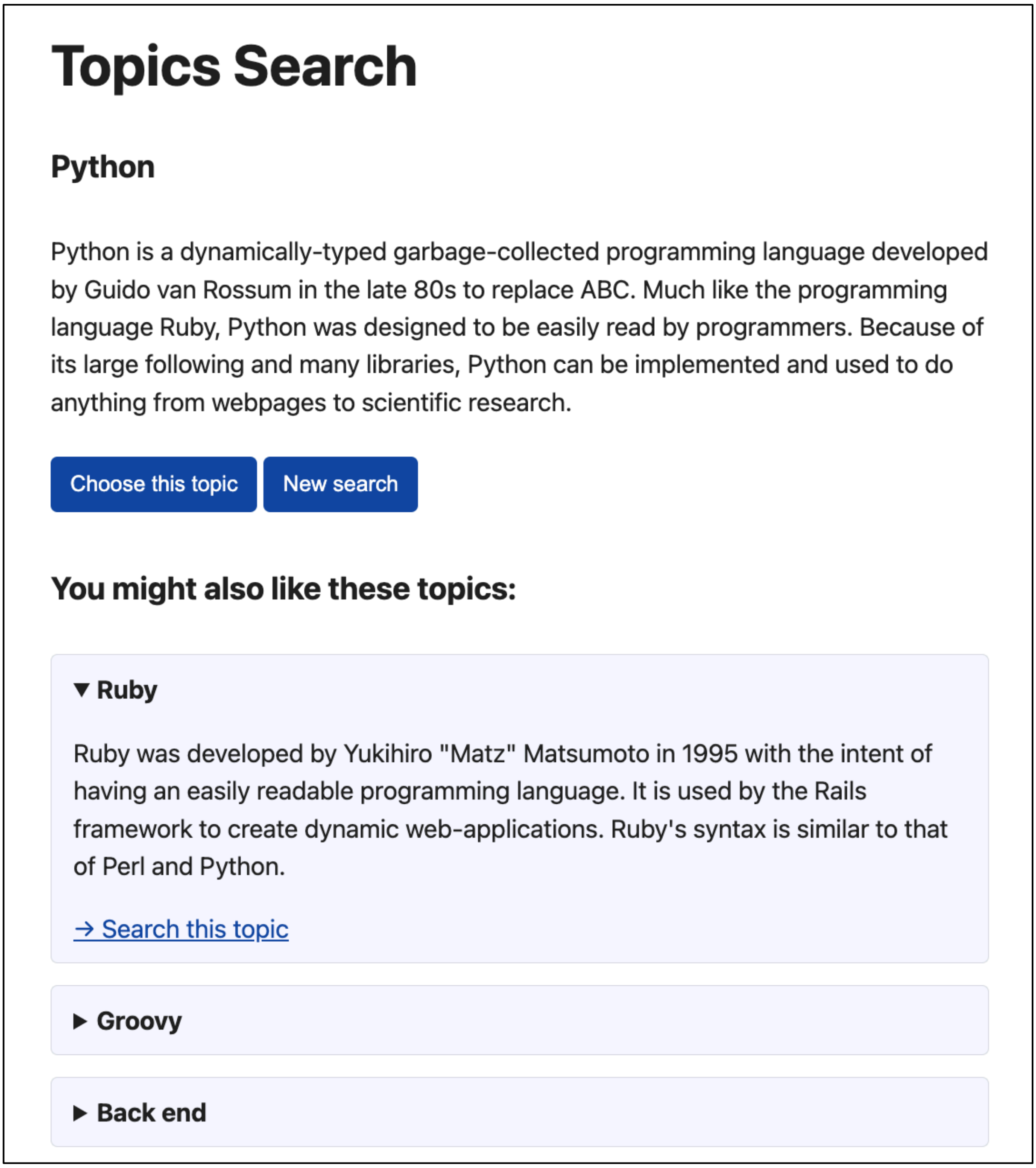
Materials

The Topics Recommender system was designed so that its recommendation algorithm could be easily configured in a laboratory setting for the purposes of this study. We explain the specific configurations of the recommendation algorithms in the relevant sections for each experiment. The main user interface integrated a search field for the names of study topics and a ranked list of topic recommendations based on the search inputs. Users began each session by entering a nickname to be used for the recording of their usage data and statistics. A description page then gave a background for the search task as well as instructions for how to use the application. The background described a fictional scenario in which the academic department was planning a new course for teaching open-source software concepts, and it asked students to identify specialized topics that they would like to learn in the course. Participants were then instructed to use the recommendation system and choose their topics of interest. As users typed the name of a search topic into the text input, an autocomplete function would show the names of existing topics in the database. Once the name of a topic was submitted, the application would then display a description of the selected topic along with a list of 10 other topics selected by the recommendation algorithm (**Figure 1**). The user can then choose to select their searched topic, read descriptions of recommended topics, perform a search from a recommended topic, or start a fresh new search.

$$\text{tfidf}(t, d) = f_{t,d} \times \log \frac{N}{|\{d_t \in D : t \in d_t\}|} \qquad \text{(Eq. 1)}$$

The data for the recommendations came from a collection of descriptive tags used to label source code repositories on GitHub.com, called GitHub Topics (Note 1). We elected to use only topics flagged as curated by the GitHub.com user community to ensure that the data included descriptive sentences for each topic. Aside from requiring user-curated topics, no other filtering or classification of the data was performed. A total of 654 topics and their community-authored descriptions were downloaded through the GitHub REST API (Note 2). After cleaning the raw data, we processed the descriptive texts to calculate similarity scores between all topics. These scores were determined by first computing tokenized word vectors with term frequency-inverse document frequency (TF-IDF) weighting and then calculating the cosine similarity between the word vectors of all the documents (i.e., topic descriptions). The TF-IDF weights for each term contained within a document were calculated as the product of term frequency within the document with the inverse of its document frequency across the entire dataset (**Eq. 1**). For term *t* in document *d*, TF-IDF is calculated from the term frequency within document $f_{t,d}$, the total number of documents *N* in dataset *D*, and the number of documents in *D* containing *t*. This approach places greater weight on terms that appear less commonly throughout the dataset and reflects the importance of terms within the context of their document (Rajaraman & Ullman, 2011).

The similarity scores resulted in a range of 0.0 to 1.0, where 1.0 represents completely identical documents. These similarity scores were calculated from the original English language descriptions and served as the basis for generating recommendations in both the English and Japanese user interfaces. Japanese descriptions of each topic were generated using machine translation provided by the DeepL API (Note 3) and verified in preliminary experiments. The Japanese interface of the Topics Recommender was subsequently used by all participants in Experiment 1 and all but one participant in Experiment 2.



Topics Search

Python

Python is a dynamically-typed garbage-collected programming language developed by Guido van Rossum in the late 80s to replace ABC. Much like the programming language Ruby, Python was designed to be easily read by programmers. Because of its large following and many libraries, Python can be implemented and used to do anything from webpages to scientific research.

You might also like these topics:

▼ **Ruby**

Ruby was developed by Yukihiro "Matz" Matsumoto in 1995 with the intent of having an easily readable programming language. It is used by the Rails framework to create dynamic web-applications. Ruby's syntax is similar to that of Perl and Python.

Figure 1

A screenshot of sample search results for the input "python" with expandable list items containing recommended topics and their descriptions

The measured effects from questionnaires in both experiments included user experiences of recommendation qualities (usefulness, novelty, accuracy, and satisfaction), user preferences for diversity in the recommendations they receive, and the NCC dimensions of decisiveness and need for predictability. The perceived recommendation qualities of usefulness, novelty, and accuracy were treated as independent factors contributing to the overall satisfaction of the user. Participants indicated their preference for recommendation diversity on a scale that placed novelty on one end and similarity on the other. Recommendation qualities were indicated by responses on a 7-point Likert scale representing the degree to which participants agreed with the given sentences (e.g., "I received new or unexpected recommendations").

The NCC items used a similar structure for responses to sentences taken from the Japanese Need for Closure Scale (J-NCS), which was translated and validated by Suzuki and Sakurai (2003) from the original Need for Closure Scale (Kurglanski & Webster, 1996). In their factor analysis of the J-NCS, the 20-item scale exhibited multidimensional factors which they labeled "decisiveness", "preference for order", and "preference for predictability". For our NCC questionnaire, we selected five items from each dimension of decisiveness and need for predictability for a total of 10 items relevant to the use of recommender systems. Since the J-NCS was first published, Roets and Van Hiel (2007) proposed an updated version of the NCS to address previously mentioned issues with items on the decisiveness subscale of the original NCS, although a Japanese version of these items was not available at the time of the current study. All the questionnaire items we selected are provided in the Appendix along with English translations created solely for the benefit of this article.

Experiment 1

The aim of the first experiment was to explore the differences in user perceptions of recommendations when they are generated with algorithms of varying diversity. The research questions guiding this experiment were as follows:

E1Q1. What are the different effects on the user from recommendations based on prediction accuracy versus those based on statistical diversity?

E1Q2. How satisfied are users with diverse recommendations compared to accurate recommendations?

E1Q3. How do users with different needs for cognitive closure evaluate various recommendations?

We configured the Topics Recommender with three distinct recommendation algorithms. The first algorithm (hereafter referred to as "TFIDF") employs the previously described similarity scores calculated with TF-IDF weighting to consistently return the top 10 topics that are most similar to the one searched by the user in descending order. The second algorithm (hereafter referred to as "semi-accurate" or "SA") returned 10 random topics from a subset of topics that had TF-IDF weighted similarity scores higher than a specified threshold. We calculated a threshold one standard deviation above the mean of similarity scores across the entire dataset and verified it in preliminary experiments. The third algorithm (hereafter referred to as "accuracy-free" or "AF") followed a similar approach to the SA algorithm, but its subset of topics included all of those with nonzero similarity scores for the searched topic. In other words, any topic that shared one or more relevant words with the searched topic was included in the AF pool for random selection. This algorithm was included as the control to which user evaluations of the TFIDF and SA algorithms may be compared.

A total of 30 students voluntarily participated in the experiment. Upon starting a new session, each participant was assigned one of the three recommendation algorithms without their knowledge, resulting in cohorts of 10 participants for each algorithm. The task included instructions to select a total of three topics for the fictional course described in the background. The system then logged their topic searches, recommendations generated by the algorithm, recommendations followed by users, and the three topics ultimately chosen by each user.

Experiment 2

The aim of the second experiment was to further explore the role of user qualities in the evaluation of recommendations. We examined participant reports of their recommendation preferences before using our system compared to after using it. Additionally, we further explored the effects of user qualities on their perceptions of recommendations generated without prediction accuracy bias. The research questions guiding this experiment were as follows:

E2Q1. How do user preferences for diversity differ after receiving recommendations compared to before using it?

E2Q2. How do users perceive the qualities of recommendations that are generated without prediction accuracy bias?

E2Q3. What influence does need for cognitive closure have on user perceptions of recommendation qualities?

We configured the Topics Recommender to use the AF algorithm from Experiment 1 for all users in order to control for the effects of prediction accuracy on the perceptions of participants. The same fictional scenario was again given as the background of this experiment, but participants were asked to choose a total of five different topics to give them greater exposure to the recommendations generated by the system.

A total of 31 students voluntarily participated in the experiment. None of the participants had any connection to the first experiment or prior experience with the Topics Recommender system. All participants answered the NCC questionnaire before using the system to control for the possibility of temporal variance in their dispositional NCC. The posttest questionnaire then included items for the recommendation qualities of usefulness, novelty, accuracy, and satisfaction. Participants were grouped by those who received the item for recommendation preference in the posttest questionnaire versus the pretest questionnaire. The Posttest cohort included a total of 16 participants, while the Pretest cohort included 15 participants. The system usage statistics of the participants were logged in the same way as in Experiment 1.

Results

In our analysis of the results, we examined the similarity scores of topics used in the recommendation algorithms as well as the ordinal data of Likert responses on the questionnaires. Since the similarity scores were normalized values, they could be analyzed directly; however, the responses to questionnaire items were first converted to rankings to account for differences in participant interpretations of the Likert scales.

The diversity of the recommendations given to each user was represented by the mean cosine distance between all items in a set of recommendations. We compared the mean diversity scores of all sets for each user with one-way ANOVA and post hoc Tukey test to determine significant differences between cohorts as this parametric test is well-suited for detecting significant differences across multiple cohorts when data are normally distributed. A similar approach was used to determine the diversity of topics chosen by each participant, with the only difference being each participant could have multiple sets of recommendations, but only a single set of chosen topics.

For questionnaire responses, we calculated statistical significance with the Kruskal–Wallis H test, as appropriate for ordinal data (Gibbons, 1993). A post hoc Dunn's test with Bonferroni correction was used in Experiment 1 to identify pairs of cohorts with significant differences. To measure effect size, we used Mann–Whitney's U test for common language effect size (McGraw & Wong, 1992), providing an interpretable measure of practical significance for differences in ordinal responses between cohorts. The relationships between questionnaire items were additionally examined using Spearman's correlation to assess monotonic trends between ranks.

Experiment 1 Results

As expected from the different designs of the algorithms, the recommendation diversity was significantly different between the three cohorts (TFIDF: $M = 0.828$, $SD = 0.020$; SA: $M = 0.922$, $SD = 0.009$; AF: $M = 0.960$, $SD = 0.007$). This was indicated by the results of a one-way ANOVA ($F(2, 27) = 265.433$; $p < .001$) with post hoc Tukey test ($\alpha = 0.05$). However, the diversity of the chosen topics did not significantly differ between cohorts (TFIDF: $M = 0.913$, $SD = 0.037$; SA: $M = 0.938$, $SD = 0.034$; AF: $M = 0.943$, $SD = 0.049$), as indicated by a one-way ANOVA ($F(2, 27) = 1.489$; $p = .244$). **Figure 2** illustrates the contrast between recommendation diversity and user choice diversity.

The results of the Kruskal–Wallis H test are shown in **Table 2** and **Table 3**. Items related to the need for cognitive closure were aggregated into a cumulative score for all 10 items as well as separate scores for decisiveness and need for predictability. Cronbach's alpha revealed that internal consistency was sufficient for each subscale (decisiveness $\alpha = .755$; need for predictability $\alpha = .702$) but poor for the cumulative scale ($\alpha = .209$).

Participants largely rated recommendation qualities positively as each quality received a median response of 5.5 or higher within each cohort as well as from all participants (usefulness *Mdn* = 6.0; novelty *Mdn* = 6.0; accuracy *Mdn* = 5.5; satisfaction *Mdn* = 6.0). The results of the Kruskal–Wallis H test showed no significant differences between the cohorts (**Table 2**).

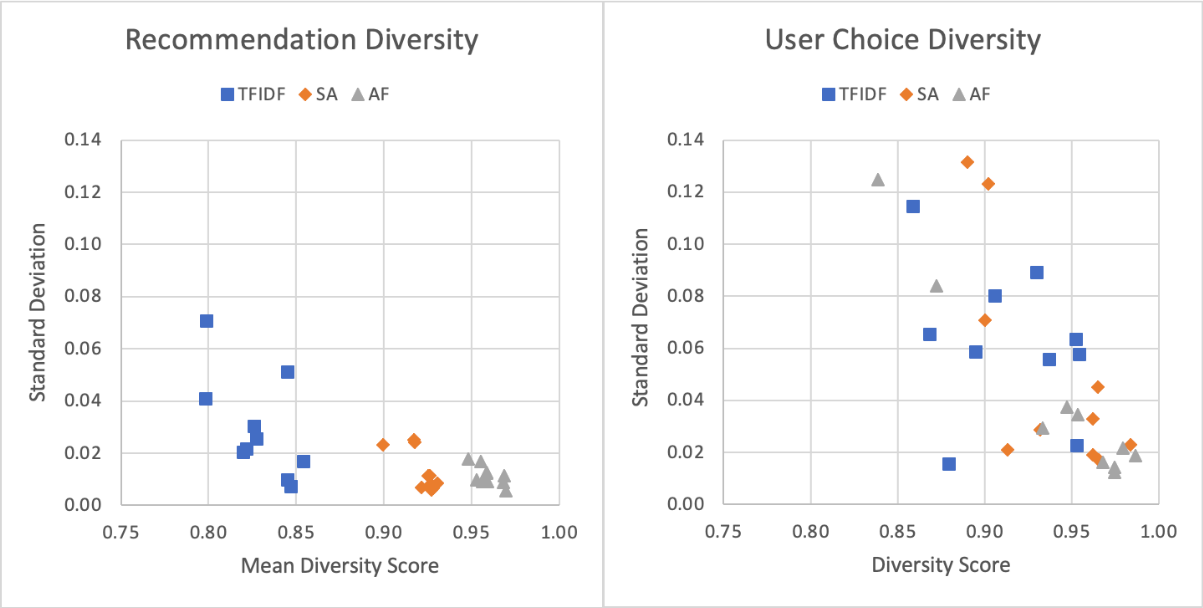


Figure 2

Differences in the diversity of recommendation lists shown to each participant (left) versus topics chosen by each participant (right) grouped by the algorithm for each participant.

Table 2

Experiment 1 questionnaire responses for items related to recommendation qualities

ERS Qualities	Usefulness		Novelty		Accuracy		Satisfaction	
	Mdn	*M* Ranks	*Mdn*	*M* Ranks	*Mdn*	*M* Ranks	*Mdn*	*M* Ranks
TFIDF	6.5	19.55	6.0	16.60	5.5	16.85	6.0	19.00
SA	6.0	15.30	4.0	11.75	6.0	17.50	6.0	16.00
AF	6.0	11.65	6.5	18.15	5.0	12.15	5.0	11.50
Kruskal-Wallis *H* $\chi^2(2)$	4.650		3.108		2.393		4.106	

Table 3

Experiment 1 questionnaire responses for items related to user qualities

User Qualities	Preference	Need for Closure[a]	Decisiveness[b]	Need for Predictability[b]

	Mdn	*M* Ranks	*Mdn*	*M* Ranks	*Mdn*	*M* Ranks	*Mdn*	*M* Ranks
TFIDF	6.0	19.30	38.0	23.40	22.0	19.95	18.0	18.50
SA	6.0	18.75	32.5	11.40	18.0	13.75	15.0	13.25
AF	4.5	8.45	33.5	11.70	13.5	12.80	17.0	14.75
Kruskal-Wallis *H* $\chi^2(2)$	10.857**		12.191**		3.925		1.899	

** p < .01

[a] Poor internal consistency (Cronbach's α = .21)

[b] Sufficient internal consistency (Cronbach's α > .70)

For items related to user qualities, the AF cohort showed a significantly lower preference for similar recommendations, as indicated by a post hoc Dunn's test with Bonferroni correction (α = .017). The common language effect size from Mann–Whitney's U test between the AF cohort and the two others was *CL* = .870 for the TFIDF cohort and *CL* = .835 for the SA cohort. This finding suggested that participants who used more accurate recommendation algorithms generally preferred similar recommendations over novel ones in the posttest questionnaire.

Table 4

Spearman's correlation between Experiment 1 questionnaire items and NCC scores among cohorts and all participants

r_s		USE	NOV	ACC	SAT	NCC	DEC	NFP
Reccomendation Preference	All	.42*	-.19	.19	.37*	.23	-.04	.34
	TFIDF	.24	-.12	.00	.00	.53	.08	.44
	SA	.48	-.57	-.03	.40	-.35	-.31	.22
	AF	.28	.41	.35	.19	.26	-.25	.53
Need for Predictability (NFP)	All	-.12	-.17	.23	-.12	.30	-.45*	
	TFIDF	-.20	-.24	-.19	-.23	.47	-.65*	
	SA	-.02	-.39	.45	.09	.03	-.61*	
	AF	-.30	-.12	.66*	-.19	.30	-.46	
Decisiveness (DEC)	All	.36	.31	-.10	.22	.68**		
	TFIDF	.57	-.15	.36	.56	.23		
	SA	-.10	.49	-.30	-.07	.73*		
	AF	.60	.50	-.48	.14	.58		
Need for Closure (NCC)	All	.33	.30	.03	.19			
	TFIDF	.48	-.16	.03	.34			
	SA	-.39	.30	.03	-.25			
	AF	.39	.50	.12	.01			
Satisfaction (SAT)	All	.80**	.07	.23				
	TFIDF	.90**	-.25	.45				

	SA	.84**	.16	-.02
	AF	.45	.76*	-.17
Accuracy (ACC)	All	.11	-.36	
	TFIDF	.33	-.25	
	SA	-.25	-.39	
	AF	-.33	-.33	
Novelty (NOV)	All	.09		
	TFIDF	-.11		
	SA	-.03		
	AF	.67*		

* *p* < .05

** *p* < .01

We also calculated Spearman's correlation between questionnaire items and NCC scores (**Table 4**). The results showed that user satisfaction correlated strongly with perceived usefulness for the TFIDF cohort, the SA cohort and all participants; however, for the AF cohort, novelty correlated with both satisfaction and usefulness. Among the user qualities, there was no significantly strong correlation between reported recommendation preferences and need for cognitive closure. While cumulative NCC correlated with decisiveness among all participants and the SA cohort in particular, all participants and cohorts confirmed a negative correlation between decisiveness and need for predictability.

The relationship between the user recommendation preferences and perceived recommendation qualities varied among the cohorts; however, among all participants, those who had a stronger preference for similar recommendations generally gave higher ratings of usefulness and satisfaction. In addition, the AF cohort responses showed a correlation between the need for predictability and the perceived accuracy of the recommendations. This implies that participants who had a stronger need for predictability reported seeing more accuracy among recommendations generated without a bias for prediction accuracy.

Interestingly, the level of diversity in the recommendations did not appear to have an effect on participant evaluations of the recommendations they received, but it was connected to their reported preferences as those who saw more accurate recommendations reported a stronger preference for similarity after using the system (E1Q1). User satisfaction was connected to these preferences for diversity and perceptions of different recommendation qualities for different cohorts. Experiences of usefulness were connected to greater satisfaction for the participants who saw relatively accurate recommendations, but those who saw accuracy-free recommendations found novelty to be more useful and satisfying (E1Q2). Moreover, need for predictability was connected to perceptions of accuracy among users who saw lists of recommendations generated without prediction accuracy (E1Q3).

Experiment 2 Results

As with the first experiment, we calculated the diversity of recommendation lists and user choices as the mean cosine distance between vectorized topic descriptions. Recommendation diversity was consistent between the Pretest cohort and Posttest cohort (Posttest: *M* = .959, *SD* = .007; Pretest: *M* = .960, *SD* = .004), as indicated by one-way ANOVA (*F*(1, 29) = .258; *p* = .616). Similarly, the analysis of user choices showed no differences in item diversity (Posttest: *M* = .951, *SD* = .020; Pretest: *M* = .961, *SD* = .018; ANOVA *F*(1, 29) = 2.263; *p* = .143).

The questionnaire items for user qualities and recommendation qualities were analyzed in the same way as in Experiment 1, and their medians, mean ranks, and Kruskal–Wallis H values are provided in **Table 5** and **Table 6**. Perceptions of recommendation qualities did not differ between the cohorts, and median ratings were generally positive across all participants (usefulness *Mdn* = 6.0; novelty *Mdn* = 5.0; accuracy *Mdn* = 6.0; satisfaction *Mdn* = 6.0).

No significant differences were observed between cohorts for the cumulative NCC scores or for decisiveness and need for predictability. In this experiment, the internal consistency of the cumulative scale was poor (Cronbach's *α* = .37), while that of the decisiveness subscale was good (*α* = .87), and need for predictability was sufficient (*α* = .77). The only difference between the cohorts was a significantly greater preference for similar recommendations in the Posttest cohort (*p* = .017) (**Table 6**) compared to the preferences of the Pretest cohort which were more balanced and leaned slightly toward novel recommendations (**Figure 3**). This is striking given their high evaluations of accuracy, although not inconsistent with our findings in Experiment 1. The common language effect size for the difference in reported preferences was *CL* = .752 according to Mann–Whitney U.

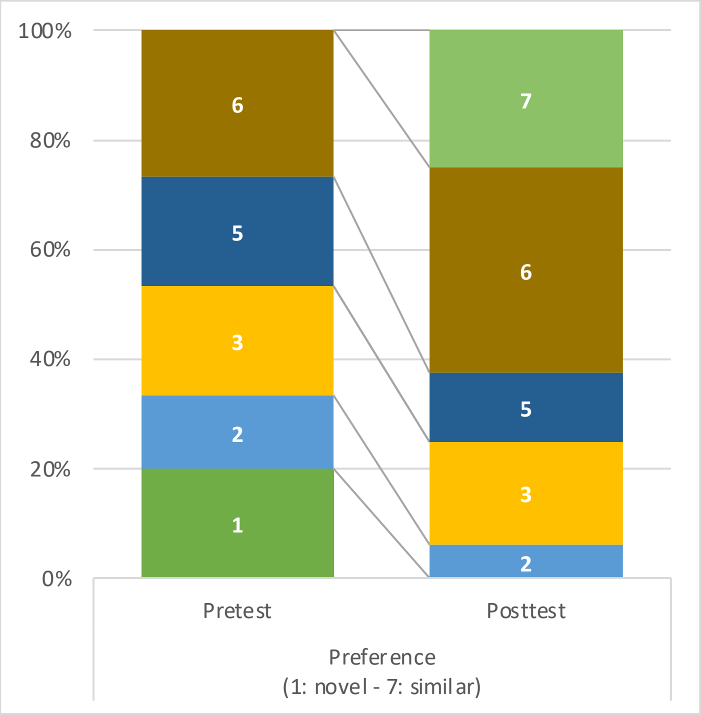


Figure 3

Distribution of Likert responses for the recommendation preferences of the Pretest and Posttest cohorts (1: novel recommendations, 7: similar recommendations)

When analyzing correlations between the questionnaire items (**Table 7**), we observed relationships between the recommendation qualities of satisfaction, perceived usefulness, and perceived accuracy among all participants and among the Pretest cohort in particular. For the Posttest cohort, only satisfaction and usefulness correlated significantly. In other words, recommendation accuracy was weighted more heavily for users who considered their preferences for recommendation diversity before using the system. This finding suggests that priming users to think about their preference for recommendation diversity may strengthen the relationships that perceived accuracy holds with perceived usefulness and satisfaction.

With regard to NCC scores, the two subscales once again correlated negatively with each other, although the effect was too weak to be significant in the Pretest cohort. Connections between perceived recommendation qualities and the need for cognitive closure were also observed. In the Posttest cohort, both satisfaction and perceived usefulness were predicted by decisiveness, while perceived usefulness correlated negatively with need for predictability. Among all participants, higher usefulness and satisfaction ratings were associated with a lower need for predictability. This suggests that participants who were less averse to ambiguity and more effective at making decisions found the recommendations to be more useful and were more satisfied with the topic recommendations overall, but this effect was weaker for those who considered their diversity preferences before performing their evaluations.

Table 5

Experiment 2 questionnaire responses for items related to recommendation qualities

ERS Qualities	Usefulness		Novelty		Accuracy		Satisfaction	
	Mdn	*M* Ranks	*Mdn*	*M* Ranks	*Mdn*	*M* Ranks	*Mdn*	*M* Ranks
Posttest	6.0	16.13	5.0	16.06	6.0	18.94	5.5	15.66
Pretest	5.0	15.87	5.0	15.93	5.0	12.87	6.0	16.37
Kruskal-Wallis *H* $\chi^2(1)$	0.007		0.002		3.637		0.050	

Table 6

Experiment 2 questionnaire responses for user qualities

User Qualities			Need for				Need for	
	Preference		Closure[a]		Decisiveness[b]		Predictability[c]	
	Mdn	*M* Ranks	*Mdn*	*M* Ranks	*Mdn*	*M* Ranks	*Mdn*	*M* Ranks

Posttest	6.0	19.78	30.5	14.06	14.0	14.59	17.0	15.44
Pretest	3.0	12.00	35.0	18.07	18.0	17.50	18.0	16.60
Kruskal-Wallis *H* $\chi^2(1)$	6.007*		1.509		0.793		0.127	

* $p < .05$

[a] Poor internal consistency (Cronbach's $\alpha = .37$)

[b] Good internal consistency (Cronbach's $\alpha = .87$)

[c] Sufficient internal consistency (Cronbach's $\alpha = .77$)

Table 7

Spearman's correlation between Experiment 2 questionnaire items and NCC scores among cohorts and all participants

r_s		USE	NOV	ACC	SAT	NCC	DEC	NFP
Reccomendation Preference	All	-.27	-.02	.18	-.23	.14	-.28	.48**
	Posttest	-.30	-.10	.27	-.33	-.05	-.35	.27
	Pretest	-.32	.02	-.17	-.20	.58*	-.19	.79**
Need for Predictability (NFP)	All	-.53**	.03	-.35	-.37*	.33	-.54**	
	Posttest	-.67**	.04	-.32	-.44	.11	-.69**	
	Pretest	-.46	.00	-.36	-.39	.50	-.48	
Decisiveness (DEC)	All	.23	.18	-.14	.22	.58**		
	Posttest	.61*	.27	.06	.72**	.62*		
	Pretest	-.20	.05	-.31	-.25	.49		
Need for Closure (NCC)	All	-.28	.17	-.45*	-.13			
	Posttest	.11	.49	-.22	.55*			
	Pretest	-.69**	-.02	-.61*	-.71**			
Satisfaction (SAT)	All	.64**	.31	.37*				
	Posttest	.54*	.48	.12				
	Pretest	.72**	.18	.63*				
Accuracy (ACC)	All	.38*	-.01					
	Posttest	.26	.03					
	Pretest	.52*	-.05					
Novelty (NOV)	All	.09						
	Posttest	-.13						
	Pretest	.30						

* $p < .05$

** $p < .01$

The results of this experiment showed that recommendation preferences can be influenced by the use of the system itself. While there was generally a wide spread of preferences in both the Posttest and the Pretest cohorts, they tended more towards similarity for the Posttest cohort compared to the more centrally balanced preferences of the Pretest cohort (E2Q1). The results also showed that users who perceived their recommendations as useful and accurate were satisfied despite the recommendations being statistically inaccurate (E2Q2). This effect may have been enhanced by first prompting the user to consider their preferences for diverse recommendations before using the system. Levels of the NCC dimensions decisiveness and need for predictability also appeared to influence perceptions of accuracy, usefulness, and user satisfaction, but this effect may have been mitigated by priming users to think about their preferences beforehand (E2Q3).

Integrated Analysis

Different recommendation preferences were observed between the cohorts in Experiment 1 and Experiment 2. When comparing all five cohorts via a single Kruskal–Wallis H test, a significant difference was observed in reported preferences ($\chi 2(4) = 17.6$, $p = .001$; Pretest *M* rank = 19.83; Posttest *M* rank = 35.78; AF *M* rank = 20.8; SA *M* rank = 40.1; and TFIDF *M* rank = 41.2), as determined by post hoc Dunn’s test (Bonferroni $\alpha = .005$), indicating that preferences of the Pretest cohort differed significantly from those of both the SA cohort and the TFIDF cohort. These results from both experiments suggest that users who experienced more accurate recommendations were more inclined to report a stronger preference for accuracy.

We further examined cohorts by combining cohorts with similar experimental properties. The AF cohort in Experiment 1 was similar to the Posttest cohort in Experiment 2 in that they both received unbiased recommendations before considering their preferences for recommendation diversity. Additionally, the TFIDF and SA cohorts in Experiment 1 also reported their recommendation preferences after using the system, although their recommendations were biased towards accuracy. Combining samples based on these traits, we analyzed questionnaire responses for cohorts of participants who reported their preferences after receiving recommendations generated without bias (AF-Posttest) as well as all participants who reported preferences in post-test questionnaires (All Posttest). These results are shown in **Table 8**, along with the combined results for all participants in Experiment 1 and Experiment 2.

Table 8

Spearman’s correlation between questionnaire items for the combined cohorts of All participants (N = 61), All Posttest (n = 46), and AF-Posttest (n = 26)

r_s		USE	NOV	ACC	SAT	NCC	DEC	NFP
Reccomendation Preference	All	.05	-.07	.22	.03	.19	.16	.40**
	All Posttest	.12	-.17	.24	.07	.11	-.13	.33*
	AF-Posttest	-.17	-.04	.34	-.15	.02	-.35	.38

Need for Predictability (NFP)	All	-.36**	-.05	-.10	-.28*	.31*	-.51**
	All Posttest	-.32*	-.07	.03	-.26	.21	-.54**
	AF-Posttest	-.53**	.04	-.01	-.36	.16	-.63**
Decisiveness (DEC)	All	-.12	-.06	.03	-.01	.19	
	All Posttest	.49**	.29	-.01	.45**	.69**	
	AF-Posttest	.61**	.31	-.08	.53**		
Need for Closure (NCC)	All	-.03	.21	-.21	-.01		
	All Posttest	.27	.35*	-.02	.33*		
	AF-Posttest	.21	.49*	-.10	.34		
Satisfaction (SAT)	All	.70**	.21	.32*			
	All Posttest	.70**	.23	.18			
	AF-Posttest	.53**	.48*	.10			
Accuracy (ACC)	All	.27*	-.17				
	All Posttest	.15	-.25				
	AF-Posttest	.16	-.17				
Novelty (NOV)	All	.09					
	All Posttest	.00					
	AF-Posttest	.10					

* *p* < .05
** *p* < .01

While user preferences were influenced by recommendation diversity, we also observed a positive correlation among all participants between need for predictability and preferences for similar recommendations. This correlation was strong in the Pretest cohort but weak among participants reporting their preferences in posttest questionnaires, suggesting that using the Topics Recommender may have reduced the influence of individual dispositions on recommendation preferences. Taking this into consideration along with the differences in reported preferences across all cohorts, it would appear that user preferences may have been initially based on dispositional need for predictability but then influenced by the algorithmic diversity or accuracy of the recommendations they received, where more similar recommendations influenced stronger preferences for accuracy.

The NCC user qualities of decisiveness and need for predictability were also linked to perceptions of recommendation qualities, and this effect was stronger for participants who received unbiased recommendations. Perceived usefulness was positively linked to decisiveness and negatively linked to need for predictability among participants in the All-Posttest cohort. This effect was stronger in the AF-Posttest cohort, in which the recommendations were not accurate by the objective measure. The negative relationship between perceived usefulness and need for predictability persisted to a weak degree for all participants but the relationship between usefulness and decisiveness did not. Interestingly, the observed effect of decisiveness on perceived usefulness is eliminated when taking the Pretest cohort's responses into account, while the effect of need for predictability stays nearly the same. This finding suggests that priming users to consider recommendation diversity before using the system likely mitigates the influence of decisiveness on perceptions of recommendation usefulness but not that of need for predictability.

Discussion

In this study, we did not find that differences in algorithmic recommendation diversity could explain differences in user perceptions of recommendation qualities. Participants consistently rated their recommendations positively for usefulness, novelty, accuracy, and satisfaction across all cohorts in both experiments despite the algorithmically more accurate recommendations seen by the TFIDF and SA cohorts of Experiment 1. This suggests that user evaluations were influenced by other factors not directly observed from the recommendations themselves. While the degree of objective diversity played a minor role, we found that psychological dispositions had a larger impact with some mitigation observed when users considered the prospect of receiving diverse recommendations in advance.

The primary indicators of user satisfaction included perceived usefulness and novelty depending on the task contexts. Perceived usefulness was strong for all participants, but it was especially important to participants in the TFIDF and SA cohorts where recommendations were more accurate. Novelty was strongest in the AF and AF-Posttest cohorts, where participants were given a greater variety of recommendations to explore without being asked to consider accuracy ahead of time. These findings reinforce the importance of aligning recommendation strategies with task contexts where accurate recommendations may be more suitable for practical tasks, while diversity and novelty enhance user satisfaction in exploratory tasks. This context-dependent approach reflects the nuanced nature of satisfaction and aligns with previous research that portrays serendipity as a combination of both novelty and usefulness (Kotkov et al., 2018; Chen et al., 2019).

The effects of psychological dispositions on usefulness and satisfaction also appeared to be context dependent. Specifically, the NCC dimensions showed stronger relationships in the AF-Posttest cohort which may be due to the greater diversity experienced by participants. The recommendations generated without prediction accuracy bias exposed participants to more unexpected items and required them to exert greater effort to process their recommendation lists. This extra cognitive load may have consequently lowered the ability of participants to perform selective processing when considering alternatives (Webster & Kruglanski, 1994). Greater effort in selective processing may also explain why those with greater need for predictability and lower decisiveness experienced fewer benefits from the recommendations. Having a lower tendency for selective processing and greater openness to the unexpected items may explain why novelty was more satisfactory for the AF-Posttest cohort. On the other hand, exposure to objectively accurate recommendations and considering accuracy ahead of time may have influenced participants' expectations and the weights they used for evaluation (Haeubl & Murray, 2003), as demonstrated by the TFIDF, SA, and Pretest cohorts.

Our findings align with learner-centered design principles and emphasize the need for ERS to adapt to different user traits, such as the NCC dimensions of decisiveness and need for predictability, to enhance satisfaction and accommodate different learning contexts. This is consistent with the notion that systems should be designed to adapt to learners' psychological and contextual differences, as described by Gronseth, Michela, & Ugwu (2020). In highlighting the influence of these psychological traits, this study reinforces theories of self-regulated learning, which advocate for adaptive systems to help learners navigate educational content autonomously.

This study supports other findings on the influences of interactive system use and priming on user traits. With regard to recommender systems where the user's task is essentially decision making, various presentation qualities are known to influence the ultimate decision in different ways (Jameson et al., 2015). The current study showed that the level of diversity also influences how users think about their decisions. With regard to evaluating options, the interaction between system properties and the greater decision-making context has been found to influence user preferences in other studies (Haeubl & Murray, 2003), although this effect could be mitigated with priming strategies that moderate preexisting attitudes toward a subject (Kim et al., 2021). We saw this in our own findings where user decisiveness was linked to perceived usefulness and satisfaction except for those who considered recommendation diversity before using the system. By guiding users to focus on specific aspects of recommendations, such priming techniques may lower cognitive load and improve decision making in educational contexts for users with a range of psychological needs. This supports learning theories that incorporate cognitive load theory in educational contexts and advocate for reducing extraneous load to enhance learning efficiency (e.g., Mayer, 2022).

Limitations and Future Work

As with any study using Likert scales, the interpretation of scale size and associated terms is a significant limitation of this study. In particular, differences in Japanese translations may have affected the interpretation of terms such as "accuracy," "novelty," and "similarity." For example, the item asking about novelty used a Japanese word meaning "unexpected," but the item asking about recommendation preference used a Japanese word meaning "new," which may have influenced how participants categorized these qualities in their minds. In addition, a study by Chen et al. (2019) revealed that experiences of novelty had a direct relationship with user satisfaction, but unexpectedness did not. If those findings are generalizable across languages and cultures, they could surely have implications for the results of this study as well.

Another limitation comes from the construction of the item asking about recommendation preference, which dichotomized the qualities of accuracy and novelty. Writing the item in this way eliminated any possibility of identifying other preferential qualities or experiences that combined aspects of novelty and accuracy, such as serendipity. Future studies should aim to develop scales that capture more granular aspects of recommendation preferences, allowing for a more detailed understanding of how users evaluate various recommendation qualities.

The size, homogeneity, and cultural aspects of the participant sample were other limitations of this study. While drawing from a single Japanese university department allowed for the control of certain confounding factors, it also limits the generalizability of the findings to broader educational contexts, particularly across different disciplines and age groups. In addition, cultural factors may influence the expression and measurement of psychological traits, warranting future studies with diverse populations to test the robustness of these findings.

The experiments were conducted in a controlled laboratory environment that does not fully capture the complexities of real-world ERS usage. Users in real-world educational settings may experience more distractions, varied time constraints, and differing levels of task engagement, which could influence their interactions with the system. Future research could include longitudinal field studies to better understand how recommendation diversity and psychological traits impact user satisfaction in more naturalistic settings.

Limitations in the Japanese translation of the Need for Closure Scale impacted scale reliability, particularly for the combined NCC score, which showed poor internal consistency. Because the J-NCS items predate the updated items for the Decisiveness dimension of the original NCS, their interpretation was likely biased toward the ability to make quick decisions rather than the need to make decisions quickly. This was evidenced by strong negative correlations between the Decisiveness and Need for Predictability subscales, consistent with the findings and criticisms of Neuberg et al. (1997). While this limited our ability to analyze need for cognitive closure as a unified construct, the subscales themselves demonstrated adequate reliability, allowing for independent comparison as originally suggested by Suzuki and Sakurai (2003). Future studies should consider using items based on the updated version of the NCS to improve internal consistency between subscales and enable a more comprehensive analysis of cognitive closure.

Conclusion

In carrying out the two experiments in this study, we explored our research questions and examined how recommendation diversity and psychological dispositions interact with user perceptions of recommendation qualities in an educational recommender system. With regards to our first question (RQ1), we found evidence that the role diversity plays on user satisfaction is highly dependent on the context. While objective diversity of recommendations had little effect on perceptions of diversity or overall evaluations, it appeared to influence the weights users put on different recommendation qualities. For example, accurate recommendations may be more suitable for practical tasks, while diverse recommendations may be more suitable for exploratory tasks. In addition, user satisfaction and evaluations were significantly influenced by dimensions of psychological need for cognitive closure (RQ2). Specifically, decisiveness contributed to user satisfaction as users found accurate recommendations to be more useful, while need for predictability was linked to a higher preference for accuracy over diversity. These effects appeared to be mitigated by priming users to think about their preferences for diversity beforehand (RQ3). Priming users to consider recommendation accuracy influenced their focus on the practicality of recommendations as their evaluations of both usefulness and satisfaction were linked to perceived accuracy. Meanwhile, those without priming placed a greater weight on novelty and diversity, indicating that priming in this manner can inherently change the user's experience with different kinds of recommendations.

Based on these findings, we propose the following considerations for practical ERS design.

1. Task-based recommendation strategies: ERS features designed for practical tasks should prioritize accuracy-based recommendations while those designed for exploratory learning should emphasize novelty and diversity.
2. Personalization based on psychological traits: User traits such as decisiveness and need for predictability should be incorporated into psychological profiles of the user in order to align recommendations with their psychological preferences and increase user satisfaction.
3. Priming as a design feature: Implementing priming techniques in the ERS interface for considering one's own preferences can guide their focus and improve satisfaction, especially for users with high levels of decisiveness or need for predictability.

These findings not only contribute to a deeper understanding of how user satisfaction is shaped by recommendation qualities and psychological traits but also provide actionable insights for improving ERS design. By tailoring recommendation strategies to specific tasks and personalizing recommendations based on users' psychological profiles, ERS can provide more satisfying and effective learning experiences. Incorporating priming techniques adds another a layer of control where systems can reduce the cognitive load of users when engaging with recommendations and dynamically adapt to their needs in real-time. Ultimately, these considerations will enable educational recommender systems to better support diverse learning goals, empower users to explore new topics with confidence, and foster more personalized, engaging, and productive learning environments.

Acknowledgements

The authors declare no conflicts of interest. The initial proposals for the experiments in this study were reviewed and approved by the Kanazawa Institute of Technology Human Subjects Research Ethics Review Committee. All participants took part on their own volition and signed written agreements allowing their data to be used anonymously for research purposes. Data privacy was ensured by associating user sessions with aliases chosen by each participant. This research was funded solely by laboratory financial resources. An artificial intelligence tool, specifically DeepL Write (Note 4), was utilized for the purpose of improving the readability of this manuscript by rephrasing portions of text.

CRediT Statement:

Robert Wesley Songer: contributed to all aspects of this research excluding resource provisioning, supervision, project management, and funding acquisition.

Tomohito Yamamoto: contributed to the conceptualization, methodology, validation, and review of the research while providing resources, supervision, project management, and funding.

References

Alhijawi, B., Awajan, A., & Fraihat, S. (2022). Survey on the Objectives of Recommender Systems: Measures, Solutions, Evaluation Methodology, and New Perspectives. *ACM Computing Surveys, 55*(5). https://doi.org/10.1145/3527449

Amer-Yahia, S., Lakshmanan, L. V. S., Vassilvitskii, S., & Yu, C. (2009). Battling Predictability and Overconcentration in Recommender Systems. *IEEE Data Engineering Bulletin, 32*(4), 33–40.

Berenbaum, H., Bredemeier, K., & Thompson, R. J. (2008). Intolerance of uncertainty: Exploring its dimensionality and associations with need for cognitive closure, psychopathology, and personality. *Journal of anxiety disorders, 22*(1), 117–125. https://doi.org/10.1016/j.janxdis.2007.01.004

Castells, P., Hurley, N. J., & Vargas, S. (2015). Novelty and Diversity in Recommender Systems. In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender Systems Handbook* (2nd ed., pp. 881-918). Springer. https://doi.org/10.1007/978-1-4899-7637-6_26

Chen, L., de Gemmis, M., Felfernig, A., Lops, P., Ricci, F., & Semeraro, G. (2013a). Human Decision Making and Recommender Systems. *ACM Transactions on Interactive Intelligent Systems*, *3*(3). https://doi.org/10.1145/2533670.2533675

Chen, L., Wu, W., & He, L. (2013b). How personality influences users' needs for recommendation diversity? *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, 829–834. https://doi.org/10.1145/2468356.2468505

Chen, L., Yang, Y., Wang, N., Yang, K., & Yuan, Q. (2019). How Serendipity Improves User Satisfaction with Recommendations? A Large-Scale User Evaluation. *The World Wide Web Conference*, 240–250. https://doi.org/10.1145/3308558.3313469

Choi, J. A., Koo, M., Choi, I., & Auh, S. (2008). Need for cognitive closure and information search strategy. *Psychology & Marketing*, *25*(11), 1027–1042. https://doi.org/10.1002/mar.20253

Erdt nee Anjorin, M., Fernandez, A., & Rensing, C. (2015). Evaluating Recommender Systems for Technology Enhanced Learning: A Quantitative Survey. *IEEE Transactions on Learning Technologies*, *8*(4), 326–344. https://doi.org/10.1109/TLT.2015.2438867

da Silva, F. L., Slodkowski, B. K., da Silva, K. K., & Cazella, S. C. (2022). A systematic literature review on educational recommender systems for teaching and learning: research trends, limitations and opportunities. *Education and Information Technologies*, 1–40. https://doi.org/10.1007/s10639-022-11341-9

Deschênes, M. (2020). Recommender systems to support learners' Agency in a Learning Context: a systematic review. *International Journal of Educational Technology in Higher Education*, *17*, 1–23. http://doi.org/10.1186/s41239-020-00219-w

Drachsler, H., Sanatos, O., & Manouselis, N. (2015). Panorama of Recommender Systems to Support Learning. In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender Systems Handbook* (2nd ed., pp. 421–451). Springer. https://doi.org/10.1007/978-1-4899-7637-

6_12

Fazeli, S., Drachsler, H., Bitter-Rijpkema, M., Brouns, F., Vegt, W. v. d., & Sloep, P. B. (2018). User-Centric Evaluation of Recommender Systems in Social Learning Platforms: Accuracy is Just the Tip of the Iceberg. *IEEE Transactions on Learning Technologies*, *11*(3), 294–306. https://doi.org/10.1109/TLT.2017.2732349

Gibbons, J. D. (1993). *Nonparametric statistics*. SAGE Publications, Inc. https://doi.org/10.4135/9781412985314

Gronseth, S. L., Michela, E., & Ugwu, L. O. (2020). Designing for diverse learners. In J. K. McDonald & R. E. West, *Design for learning: Principles, processes, and praxis*. EdTech Books. Retrieved October 20, 2024, from https://edtechbooks.org/id/designing_for_diverse_learners

Gunawardana, A., & Shani, G. (2015). Evaluating Recommender Systems. In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender Systems Handbook* (2nd ed., pp. 265–308). Springer. https://doi.org/10.1007/978-1-4899-7637-6_8

Haeubl, G., & Murray, K. (2003). Preference construction and persistence in digital marketplace: The role of electronic recommendation agents. *Journal of Consumer Psychology*, *13*, 75–91. https://doi.org/10.1207/S15327663JCP13-1&2_07

Hitsuwari, J., & Nomura, M. (2021). Developing and Validating a Japanese Version of the Multidimensional Attitude toward Ambiguity Scale (MAAS). *Psychology*, *12*(4). https://doi.org/10.4236/psych.2021.124030

Jameson, A., Willemsen, M. C., Felfernig, A., de Gemmis, M., Lops, P., Semeraro, G., & Chen, L. (2015). Human Decision Making and Recommender Systems. In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender Systems Handbook* (pp. 611–648). Springer. https://doi.org/10.1007/978-1-4899-7637-6_18

Kaminskas, M., & Bridge, D. (2016). Diversity, Serendipity, Novelty, and Coverage: A Survey and Empirical Analysis of Beyond-Accuracy Objectives in Recommender Systems. *ACM Transactions on Interactive Intelligent Systems*, *7*(1). https://doi.org/10.1145/2926720

Kapoor, K., Kumar, V., Terveen, L., Konstan, J. A., & Schrater, P. (2015). "I like to explore sometimes": Adapting to Dynamic User Novelty Preferences. *Proceedings of the 9th ACM Conference on Recommender Systems*, 19–26. https://doi.org/10.1145/2792838.2800172

Kardes, F. R., Cronley, M. L., Kellaris, J. J., & Posavac, S. S. (2004) The Role of Selective Information Processing in Price-Quality Inference. *Journal of Consumer Research*, *31*(2), 368–374. https://doi.org/10.1086/422115

Kim, E. J., Tanford, S., & Book, L. A. (2021). The Effect of Priming and Customer Reviews on Sustainable Travel Behaviors. *Journal of Travel Research*, *60*(1), 86–101. https://doi.org/10.1177/0047287519894069

Kotkov, D., Konstan, J. A., Zhao, Q., & Veijalainen, J. (2018). Investigating serendipity in recommender systems based on real user feedback. *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, 1341–1350. https://doi.org/10.1145/3167132.3167276

Kruglanski, A. W., & Webster, D. M. (1996). Motivated closing of the mind: "Seizing" and "freezing". *Psychological Review*, *103*(2), 263–283. https://doi.org/10.1037/0033-295X.103.2.263

Kruglanski, A. W., Webster, D. M., & Klem, A. (1993). Motivated resistance and openness to persuasion in the presence or absence of prior information. *Journal of Personality and*

Social Psychology, *65*(5), 861–876. https://doi.org/10.1037/0022-3514.65.5.861

Litman, J. A. (2010). Relationships between measures of I- and D-type curiosity, ambiguity tolerance, and need for closure: An initial test of the wanting-liking model of information-seeking. *Personality and Individual Differences*, *48*(4), 397–402. https://doi.org/10.1016/j.paid.2009.11.005

Lu, W., Ioannidis, S., Bhagat, S., & Lakshmanan, L. V. S. (2014). Optimal recommendations under attraction, aversion, and social influence. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 811–820. https://doi.org/10.1145/2623330.2623744

Lutz, C., Hoffmann, C., & Meckel, M. (2017). Online Serendipity: A Contextual Differentiation of Antecedents and Outcomes. *Journal of the Association for Information Science and Technology*, *68*(7). 1698–1710. https://doi.org/10.1002/asi.23771

Maccatrozzo, V., Terstall, M., Aroyo, L., & Schreiber, G. (2017). SIRUP: Serendipity In Recommendations via User Perceptions. *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, 35–44. https://doi.org/10.1145/3025171.3025185

Manouselis, N., Drachsler, H., Verbert, K., & Duval, E. (2012). *Recommender Systems for Learning*. Springer.

Marante, Y., Silva, V., Gomes Jr., J., Vitor, M., Martins, A., & De Souza, J. (2020). Evaluating Educational Recommendation Systems: a systematic mapping. *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, 912–921. Porto Alegre: SBC. https://doi.org/10.5753/cbie.sbie.2020.912

Matt, C., Benlian, A., Hess, T., & Weiß, C. (2014). Escaping from the Filter Bubble? The Effects of Novelty and Serendipity on Users' Evaluations of Online Recommendations, *ICIS 2014 Proceedings*, *67*.

Mayer, R. E. (2022). Cognitive theory of multimedia learning. In R. E. Mayer & L. Fiorella (Eds.), *The Cambridge handbook of multimedia learning* (3rd ed., pp. 57–72). Cambridge University Press. https://doi.org/10.1017/9781108894333.008

McGraw, K. O., & Wong, S. P. A common language effect size statistic. *Psychological Bulletin*, *111*(2), 361–365. https://doi.org/10.1037/0033-2909.111.2.361

McNee, S. M., Riedl, J., & Konstan, J. A. (2006). Being accurate is not enough: how accuracy metrics have hurt recommender systems. *CHI '06 Extended Abstracts on Human Factors in Computing Systems*, 1097-1101. https://doi.org/10.1145/1125451.1125659

Mehrotra, R., Shah, C., & Carterette, B. (2020). Investigating Listeners' Responses to Divergent Recommendations. *Proceedings of the 14th ACM Conference on Recommender Systems*, 692–696. https://doi.org/10.1145/3383313.3418482

Neuberg, S. L., Judice, T. N., & West, S. G. (1997). What the Need for Closure Scale measures and what it does not: Toward differentiating among related epistemic motives. *Journal of Personality and Social Psychology*, *72*(6), 1396–1412. https://doi.org/10.1037/0022-3514.72.6.1396

Pariser, E. (2011). *The filter bubble: what the Internet is hiding from you*. Penguin Press.

Pu, P., Chen, L., & Hu, R. (2011). A user-centric evaluation framework for recommender systems. *Proceedings of the 5th ACM Conference on Recommender Systems*, 157–164. https://doi.org/10.1145/2043932.2043962

Pu, P., Chen, L., & Hu, R. (2012). Evaluating recommender systems from the user's perspective: survey of the state of the art. *User Modeling and User-Adapted Interaction*, *22*, 317–355. https://doi.org/10.1007/s11257-011-9115-7

Rajaraman, A., & Ullman, J. (2011). *Data Mining. In Mining of Massive Datasets* (pp. 1-17). Cambridge University Press. https://doi.org/10.1017/CBO9781139058452.002

Roets, A., & Van Hiel, A. (2007). Separating ability from need: Clarifying the dimensional structure of the need for closure scale. *Personality and Social Psychology Bulletin*, *33*(2), 266-280. https://doi.org/10.1177/0146167206294744

Said, A., Fields, B., Jain, B. J., & Albayrak, S. (2013). User-centric Evaluation of a K-furthest Neighbor Collaborative Filtering Recommender Algorithm. *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, 1399–1408. https://doi.org/10.1145/2441776.2441933

Suzuki, K., & Sakurai, S. (2001). Ninchiteki Kanketsu Yokkyuu: Jyouhoushori Ryou Oyobi Taijinkankei kara no Kentou [The cognitive need for closure: A study in terms of information-processing quantity and social relationships]. *Tsukuba Psychological Research*, *23*, 153-160.

Suzuki, K., & Sakurai, S. (2003). Ninchiteki Kanketsu Yokkyuu Shakudo no Sakusei to Shinraisei Datousei no Kentou [The construction, reliability and validity of a Japanese Need for Closure Scale]. *The Japanese Journal of Psychology*, *74*(3), 270-275. https://doi.org/10.4992/jjpsy.74.270

Urdaneta-Ponte, M. C., Mendez-Zorrilla, A., & Oleagordia-Ruiz, I. (2021). Recommendation Systems for Education: Systematic Review. *Electronics*, *10*(14). https://doi.org/10.3390/electronics10141611

Webster, D. M., & Kruglanski, A. W. (1994). Individual differences in need for cognitive closure. *Journal of Personality and Social Psychology*, *67*(6), 1049–1062. https://doi.org/10.1037//0022-3514.67.6.1049

Wieland, M., Nordheim, G.V., & Kleinen-von Königslöw, K. (2021). One Recommender Fits All? An Exploration of User Satisfaction With Text-Based News Recommender Systems. *Media and Communication*, *9*(4). https://doi.org/10.17645/mac.v9i4.4241

Wu, W., Chen, L., & He, L.. (2013). Using personality to adjust diversity in recommender systems. *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, 225–229. https://doi.org/10.1145/2481492.2481521

Zhao, P. & Lee, D. L. (2016). How Much Novelty is Relevant? It Depends on Your Curiosity. *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 315–324. https://doi.org/10.1145/2911451.2911488

Zhou, T., Kuscsik, Z., Liu, J., Medo, M., Wakeling, J.R., & Zhang, Y. (2010). Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences*, *107*(10), 4511–4515. https://doi.org/10.1073/pnas.1000488107

Notes

Note 1. GitHub Topics, https://github.com/topics

Note 2. GitHub REST API, https://docs.github.com/en/rest

Note 3. DeepL API, https://www.deepl.com/pro-api

Note 4. DeepL Write, https://www.deepl.com/en/write

Appendix A

Questionnaire Items

Need for Closure Scale

The following items from the Japanese NCS (Suzuki & Sakurai, 2003) are provided here with our own English translations. Items with reverse wording are indicated with (R). The subscale to which each item belongs is indicated as DEC (decisiveness) or NFP (need for predictability).

1. 何が起こるか分からない状況にいるとウキウキする．(R) (NFP)
   I am excited when I'm in a situation where I don't know what is going to happen.
2. すぐに決断をしなければならない状況は不安になる．(R) (DEC)
   I find situations where I need to make a quick decision unsettling.
3. 何が起こるか分からないところには行きたくない．(NFP)
   I don't want to go where I don't know what's going to happen.
4. 決断するのにいつも苦労する．(R) (DEC)
   I always have a hard time making decisions.
5. 何が起こるか分からないような新しい状況に飛び込んでいくことは、おもしろいと思う．(R) (NFP)
   I think it's interesting to jump into new situations where you don't know what's going to happen.
6. 予想もつかないようなことをする友だちが好きだ．(R) (NFP)
   I like friends who do unpredictable things.
7. 自分は決断力がないと思う．(R) (DEC)
   I don't think I'm decisive.
8. 決めなければならないことは長い間引き延ばさず、すぐに決める方だ．(DEC)
   I'm one who decides what needs to be decided quickly, without procrastinating for too long.
9. 思いもよらないことをしそうな人と一緒にいるのは好まない．(NFP)
   I don't like to be around people who might do something I wouldn't expect.
10. 重要な決定は、たいてい素早く、自信を持って行う．(DEC)
    I usually make important decisions quickly and confidently.

Recommendation Qualities

1. 興味のあるトピックを見つけるために、推薦が役に立った。
   The recommendations were useful for finding topics of interest.
2. 新しい、あるいは思いがけない推薦をもらった。
   I received new or unexpected recommendations.
3. 検索したトピックに近いものを的確に推薦してくれた。
   I received accurate recommendations that were close to the topics I searched for.
4. タスクを完了するために、システムが推薦してくれたトピックに満足している。
   I am satisfied with the topics recommended by the system to complete the task.

Recommendation Diversity Preference

あなたは普通に推薦するアプリを使う時、自分の選んだコンテンツに対してより新しい推薦と、より類似した推薦のどちらが好みですか?

When you normally use an app that makes recommendations, do you prefer newer recommendations or more similar recommendations for your chosen content?