About the Quality of a Course Recommender System as Perceived by Students

Kerstin Wagner[1][a], Agathe Merceron[1][b], Petra Sauer[1] and Niels Pinkwart[2][c]

[1]*Berliner Hochschule für Technik, Berlin, Germany*

[2]*Deutsches Forschungszentrum für Künstliche Intelligenz, Berlin, Germany*

Keywords: Course Recommender System, Survey, Mann-Whitney U Test, Wilcoxon Signed-Rank Test, Benjamini-Hochberg Procedure.

Abstract: In this work, we present a survey of a course recommender conducted among students and its results. The course recommender system, published in our previours work (Wagner et al., 2023), is based on the nearest neighbors algorithm and aims to support students in their course enrollment; it targets above all students who did not pass all mandatory courses as indicated in the study handbook in their first or second semester at university. The primary objective of the survey was to evaluate the perceived quality of explanations and recommendations based on two presentation variants (a ranked list of courses and a set of courses), as well as the general trust in such systems. The survey included quantitative measures and demographic information from the students, so that different subgroups could be evaluated. The results indicate that students tend to trust recommender systems and that they tend to understand the explanations. No clear winner emerges between the presentation of the courses as a set and as a ranked list. The survey data explorations are available at: https://kwbln.github.io/csedu24.

1 INTRODUCTION

In the field of higher education, various recommender systems have been proposed for different purposes. According to Urdaneta-Ponte et al. (2021), course recommendations have emerged as the second most prevalent research area with 33 studies conducted on this topic. Among the articles analyzed by the authors, 25 specifically targeted students. In this work, we consider the course recommender system proposed in our previous work (Wagner et al., 2023). Our system aims to support students in their course enrollment and to help, above all, students who did not pass all mandatory courses as indicated in the study handbook in their first or second semester at university. In some contexts, like in German higher education, when enrolling in courses for their second or third semester, these students must decide whether they should repeat courses they did not pass, whether they should add new courses to their enrollment list, how many, and which ones. Our system recommends to a student *st* courses based on the courses passed by *st*'s neighbors (Wagner et al., 2023). A neighbor of the student *st* is a student who has already graduated and in the first or second semester passed courses similar to those *st* passed with grades similar to those obtained by *st*. The system recommends to the student *st* the set of courses that the majority of the nearest neighbors have passed. Let st_1 be a student who passes all courses as given in the study handbook. The evaluation of the recommended courses system with historical data shows that, on average, our system recommends to st_1 the same set of courses that st_1 has enrolled. Let st_2 be a student who failed courses in the first or second semester. The evaluation of the number of recommended courses shows that it recommends on average a smaller set of courses and different courses than st_2 enrolled in. With the assumption that st_2 is able to pass all the courses in this smaller set, the evaluation of the predicted dropout risk indicated that such a system can reduces the risk of students dropping out.

Following a user-centered approach, we conducted a survey among current students to present them the recommender system and gather their perceptions and opinions. Our aim was to address the fol-

[a] https://orcid.org/0000-0002-6182-2142
[b] https://orcid.org/0000-0003-1015-5359
[c] https://orcid.org/0000-0001-7076-9737

Wagner, K., Merceron, A., Sauer, P. and Pinkwart, N.
About the Quality of a Course Recommender System as Perceived by Students.
DOI: 10.5220/0012634900003693

In *Proceedings of the 16th International Conference on Computer Supported Education (CSEDU 2024) - Volume 2*, pages 238-246
ISBN: 978-989-758-697-2; ISSN: 2184-5026

lowing research questions: What is the level of trust that students have in course recommendation systems? How do students evaluate the quality of explanations and recommendations provided by the course recommender system in this study? To explore if the perceived quality of recommendations varies based on how the recommended courses are presented, we presented the course recommendations in two different ways: a) as a ranked list of courses sorted by their probability of being passed, and b) as a set of courses that are expected to be passed. A course is added to the list if at least one neighbor has passed it, rather than requiring a majority of neighbors to have passed the course. This approach provides students with a wider range of course options to choose from.

The paper is organized as follows. The next section provides an overview of related research. In the third section, we describe the methodology of the survey. In Section 4, the results and discussion are presented. The final section concludes the paper and discusses limitations and future directions.

2 RELATED WORK

Urdaneta-Ponte et al. (2021) provided an overview of recommendation systems for education, the education types for which they were developed, the elements they recommend, their developmental approach and implemented platforms, as well as the quality metrics to evaluate the recommendation systems. Even though studies use the same basic metrics, such as recall, there are still differences in the data basis on which they apply the metric. In some studies the recommendation system is evaluated based on a fixed number of recommended courses, Top5 or Top10 for example, in other studies the number of courses actually taken by the students is employed.

Elbadrawy and Karypis (2016) examined in their study, how various student and course groupings influence the ability to predict grades and recommend courses. The authors presented their findings by comparing the results of five recommended courses with the courses that the students had actually taken. Pardos et al. (2019) showcased methods for data synthesis to balance users' preferences and assist in decision making and evaluated the recommendations using ten recommended courses since students enrolled in between four and nine courses on average. Pardos and Jiang (2020) aimed to recommend courses "that are novel or unexpected to the student but still relevant to their interests" and recommended ten courses based on a course chosen by a student.

Other authors do not set a fixed number of recommended courses for all students. Instead, they limit the number of recommended courses to match the number of courses each student has taken. Morsy and Karypis (2019) analyzed their approaches to recommend courses in terms of their impact on students' grades and distinguished between good and bad courses to recommend good courses only. Polyzou et al. (2019) provided an interpretable framework based on students' enrollments and evaluated the recommendations for different study programs and with different characteristics. Ma et al. (2020) developed a hybrid recommender system that integrates the aspects of interest, grades, and time. Khan and Polyzou (2023) used session-based techniques to recommend courses and evaluated their suitability from a co-enrollment perspective. All compared the number of recommended courses with the courses that the students had actually taken without noting whether the courses have been passed.

Our recommender system determines the number of recommended courses based on their probability of being passed (Wagner et al., 2023). That means, we did not used a fixed number or the total number of courses taken by students. The evaluation was conducted with respect to the courses that were actually passed, which may be a smaller number compared to the courses enrolled in, as students have the option to not take exams or may fail a course.

Apart from evaluating the recommendation based on historical data, a user survey was conducted for two of the course recommendation systems that have been previously introduced. Pardos et al. (2019) ran a usability study among 20 students to analyze the alignment of the recommendation with "users' needs and to collect feedback." Pardos and Jiang (2020) had the algorithms used evaluated by 70 students as part of a survey "(1) in terms of their unexpectedness (2) successfulness / interest in taking the course (3) novelty (4) diversity of the results."

Our Contribution. In this study, a survey was conducted among students to compare how they perceive the quality of two presentations of course recommendations, a ranked list of courses versus a set of courses, which is a unique aspect of this research. Additionally, the students were asked to rate the quality of the explanations provided by the recommender system and their overall trust in such systems. Another distinctive feature of this study is the analysis of survey responses based on different subgroups of students. The results indicate that students tend to trust course recommendation systems. Furthermore, they have a positive perception of the quality of the explanations and recommendations that we provided in the survey regarding the course recommendation system. However, there is a statistically significant difference observed in one subgroup of: students who have considered dropping out of their studies. Apart from this subgroup, no significant differences were found between the other subgroups of students. The results further indicate that both presentation variants,a ranked list and a set of courses, can be equally effective, as students did not clearly favor one presentation variant over the other. Considering the statistical significance of the differences, it could be important to take into account two specific subgroups of students—those whose parents have already studied and those who have thought about dropping out of their studies—when selecting a presentation format for recommendations in recommender systems.

3 METHODOLOGY

The main objective of the project is to assist students in their course enrollments, with the main focus on students who do not study according to the plan.

The survey was carried out in two study programs, ”Architecture” (AR) and ”Computer Science and Media” (CM). Given that our research is focused on early dropout and our recommendation system is based on past academic achievements—specifically, students need to have finished a minimum of one semester—we concentrated on the second semester to choose two courses, one from each academic program, for conducting the survey during the corresponding on-site classes. As students have the flexibility to enroll in courses that are not planned for their current semester, students from the first semester or beyond are allowed to enroll in the selected courses. Two sample cases, one for each study program, were created using authentic and plausible scenarios to familiarize students with the recommendation system.

The participants were provided with an overview of the project, its objectives, and the current state of the recommender system. Students were given the option to complete the survey either through a provided link or on paper. To reach additional students, the survey was additionally distributed by email. The survey was carried out in German.

3.1 Questionnaires

The primary objective of the survey was to evaluate the perceived quality of explanations and recommendations based on two presentation formats (list and set), as well as the general trust in recommender systems. The survey included quantitative measures and demographic information from the students, so that different subgroups could be evaluated. Open-ended questions were also included to gather qualitative feedback. Participation in the survey, including providing ratings, free text responses, and demographic data, was voluntary.

We adapted relevant items from a previous study conducted by Hernandez-Bocanegra and Ziegler (2023) to suit our research questions and specific context. The adapted items were rated on a 5-point Likert scale, ranging from strong disagreement (1) to strong agreement (5). In the following, we provide the investigated categories with their items:

Perceived Explanation Quality EQ

EQ01: The explanations make me confident that I will pass the recommended courses.

EQ02: The explanations make the recommendation process clear to me.

EQ03: The explanations are convincing.

EQ04: The explanations are easy to understand.

EQ05: The explanations provide enough information for me to choose courses.

EQ06: It is clear to me what kind of data the recommendation system uses to generate recommendations.

Perceived Recommendation Quality RQ

RQ01: I understand why the courses were recommended to me.

RQ02: I can see how well the recommendations match my situation.

RQ03: I would recommend the recommendation system to others.

RQ04: I could make better decisions using the recommendation system.

General Trust in Recommender Systems GT

GT01: I would feel comfortable depending on the information from a recommendation system.

GT02: I would be confident in enrolling in the courses recommended to me by a recommender system.

GT03: I would be willing to share my past course results with a recommender system so it could recommend appropriate courses.

The survey followed a specific order, starting with obtaining consent to participate, followed by rating the explanation quality, recommendation quality, and our system. Participants also rated their general trust in recommender systems, provided demographic information, and rated the overall survey quality. The ratings for the perceived recommendation quality included the same four items for both variants, the list (RQL) and set (RQS). The order in which the list or set was rated first was randomized.

3.2 Texts and Examples Used

In the following, we provide the texts used to explain the recommender system and the generation of the course recommendations. As an example, we selected a student with good grades in their first semester but who was also not enrolled in a mandatory course during that time. In the survey, we gave the information about their academic performance in the first semester, that is, the exact grades achieved (Table 1).

Explanation of Our Recommender System. *"Our course recommendation system is based on artificial intelligence and uses the nearest-neighbor algorithm. It [the nearest-neighbor algorithm] is based on similarities between people or things. Let us say you want to have a movie night and ask your friends for movie recommendations. Your friends who have movie tastes similar to yours can give you the best recommendations. For course recommendations, the system uses the similarity of the students, whereby the similarity is calculated based only on previous performance. This means that students similar to you have passed similar courses and received similar grades in the first semester. Demographic information such as gender and age is not included. Students similar to you are therefore your neighbors. Only students who have already completed their studies are considered neighbors, that is, no students who have dropped out. Therefore, course recommendations are based on successful students."*

List of Courses. We provide below the text describing how the recommendations and the list of recommended courses for the 2nd semester of the sample student in the CM study program were generated.
"If at least 1 out of 5 neighbors have passed a course in the 2nd semester, this course is recommended for enrollment. The courses are sorted: the higher they are on the list, the more neighbors have passed them. If courses have been passed by the same number of neighbors, they are sorted according to their ID:

1. *06 Mathematics II*
2. *09 Programming II*
3. *10 Operating Systems*
4. *07 Algorithms and Data Structures*
5. *08 Database Systems*
6. *03 Fundamentals of Media Design*
7. *17 Web Engineering I"*

Table 1: 1st semester courses and the results of the example student of study program CM. Grades range from 1.0 to 5.0 with a grading scale of [1.0, 1.3, 1.7, 2.0, 2.3, 2.7, 3.0, 3.3, 3.7, 4.0, 5.0] with 1.0 being the best grade, 4.0 being the worst (just passed), and 5.0 means fail. CS = Computer Science.

ID	Course Name	Result
01	Mathematics I	2.0
02	Fundamentals of Theoretical CS	2.0
03	Fundamentals of Media Design	not enrolled
04	Technical Fundamentals of CS	1.7
05	Programming I	1.0

Set of Courses. We provide below the text describing how the recommendations and the set of recommended courses for the 2nd semester of the sample student in the CM study program were generated.
"If at least 3 out of 5 neighbors, that is, the majority of your neighbors, have passed a course in the 2nd semester, this course is recommended for enrollment. It is assumed that all courses can be passed. The courses are sorted according to their ID:

- *06 Mathematics II*
- *07 Algorithms and Data Structures*
- *09 Programming II*
- *10 Operating Systems"*

3.3 Participants

The link to the questionnaire was clicked 169 times in total. This includes both the paper questionnaires that were transferred to the survey system and the instructors and other people who received the survey for information purposes. For the current work, we filtered student questionnaires, performed quality checks, and explored the demographic data.

Filtering and Quality Checks. We only considered questionnaires in which participants have clicked at least up to the "General Trust in Recommender Systems", that is, before the demographic information, regardless of how many items they have rated, and ended up with 116 valid questionnaires from students. We filtered students in terms of the survey quality questions. 102 of 116 students positively answered the question "Can we use your data in an anonymous form for scientific purposes?". We removed the questionnaire data from one student of each study program, who answered the question "Did you perform all tasks as asked in the respective instructions?" with "I often clicked on something so I could finish

Table 2: Summary of the demographic variables and the data provided by the students in absolute and relative quantities.

Variable		Choices	Numbers	Percentage
DD01	Please indicate your semester of study.	2 n2 = not 2	68 30	69.4% 30.6%
DD02	Please state your gender.	1 = male 2 = female	51 44	53.7% 46.3%
DD03	Have your parents or one parent in your family already studied?	1 = no 2 = yes	45 51	46.9% 53.1%
DD04	Do you or at least one of your parents not possess German citizenship at birth?	1 = no 2 = yes	48 49	49.5% 50.5%
DD05	Are you taking courses as scheduled in the curriculum?	1 = no 2 = yes	40 57	41.2% 58.8%
DD06	Have you been able to take courses for credit?	1 = no 2 = yes	52 22	70.3% 29.7%
DD07	Have you ever thought about dropping out of your studies?	1 = no 2 = yes	56 38	59.6% 40.4%

quickly.” Finally, 100 questionnaires were included in the evaluation (AR: 55, CM: 45).

Demographics. Concerning demographic factors, our objective was to investigate whether certain subgroups, which may be more relevant to the students being targeted by the recommender system, particularly those who do not follow the optimal study plan outlined in the handbook and/or are at risk of dropping out from their studies, have a different perception of the recommender system compared to other students. In the following, we give the rational to introduce the demographic questions (DD) shown in Table 2. Table 2 also provides the numbers of the students who answered the questions.

DD01 Semester. While we have selected courses intended for the second semester, enrollment is also open to students from other semesters. Second-semester students are at an early stage of their academic journey and face a higher risk of dropping out from their studies compared to those in higher semesters. Non-second-semester students deviating from the study schedule may indicate academic challenges they are encountering.

DD02 Gender. From 2013 to 2022, the rate of failing the final exam in higher education in Germany has been higher among male students than among female students: On average, 4.9% male students and 2.4% female students failed (Statistisches Bundesamt, 2023). In our survey, only one person chose the ”diverse gender”. Due to concerns regarding data protection, we decided not to include this information in the study as it could potentially lead to the identification of the person and their statements.

DD03 Education background. Students whose parents did not study are underrepresented in German higher education and face special needs (Miethe et al., 2014).

DD04 Migration background. Students with a migration background are underrepresented in German higher education and drop out of their studies more often (Berthold et al., 2012).

DD05 Studying according to the curriculum. We introduce this question because our recommender system primarily focuses on students who do not study according to the study handbook.

DD06 Previous course credits. If students receive credit for courses from previous studies, they do not study according to the plan, as they skip these courses and can enroll in courses from higher semesters. These students may encounter difficulties because they do not study according to the plan given in the study handbook. However, it remains unclear whether they dropped out or not from the other program and whether they are particularly motivated and benefit from their previous experience. The limited response of only 74 students to this question may be attributed to a lack of awareness among students regarding the possibility of earning credits for courses.

DD07 Dropout thoughts. Regardless of the possible factors that may lead to student dropout, the objective of the recommender system presented in our previous work is to decrease the dropout rate (Wagner et al., 2023). In this regard, we are particularly interested in the answers of students who have already considered dropping out.

3.4 Evaluation

We aggregated the ratings given to items by students to obtain a score by category. Therefore, we handle the 5-point Likert scale as ordinal-scaled values. First, we calculated the median of the item ratings for each student and each category, such as explanation quality EQ. For instance, the median score for EQ is determined by calculating the median of the item ratings from EQ01 to EQ06. For the evaluation of the categories based on all students or within subpopulations, we aggregated the students' scores into the median of medians.

Statistical Testing for Rating Differences. We employed either the Mann-Withney U test or the Wilcoxon signed-rank test to assess the statistical significance of differences. The Mann-Withney U test evaluates unpaired data, in our case differences in the ratings of one category between supopulations. The Wilcoxon signed-rank test evaluates paired data, in our case differences in the ratings of the recommendation quality of the list RQL and the set RQS within supopulations. The significance level was set at 0.05. It is important to note that all p-values were adjusted using the Benjamini-Hochberg procedure to account for multiple testing (Matayoshi and Karumbaiah, 2021) using a false discovery rate of 0.2 and the Python package statsmodels (https://www.statsmodels.org).

4 RESULTS AND DISCUSSION

To examine the survey data, we initially assess the overall ratings of all categories among all students. Next, we examine the ratings of all categories among different subpopulations. Lastly, we compare the scores of the perceived quality of recommendation between the two presented variants: the list and the set.

4.1 General Evaluation

We present the general trend for each category, including the number of students who tend to disagree (rated the category with a score less than 3), the number of students who tend to agree (rated the category with a score higher than 3), and the number of students who were undecided (rated the category with a score of 3). Further, we describe the range and distribution of the categories' scores including their median, the lower and the upper quartiles, and outliers if applicable (Figure 1).

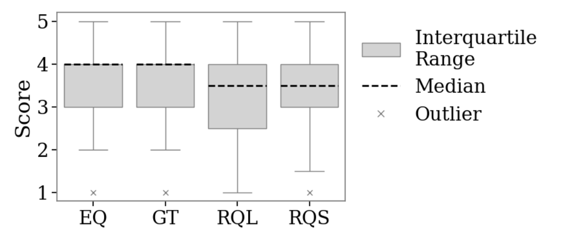

Figure 1: Distribution of the scores of all categories including all students as box plots.

The mode, which is the value with the highest frequency, for each category is "4=rather agree." The minimum for each category is "1=strongly disagree" and the maximum is "5=strongly agree." The general trends for each category are as follows.

- GT: Out of a total of 99 students who rated GT, 61 students tend to agree, 12 students tend disagree.
- EQ: Out of 100 students, 70 tend to agree, 18 students tend to disagree.
- RQ of the list (RQL): Out of 100 students, 53 tend to agree, 31 students tend to disagree.
- RQ of the set (RQS): Out of 99 students, 59 tend to agree, 22 students tend to disagree.

The students' scores for each category range from 1.0 to 5.0 with slightly different distributions (Figure 1). The perceived explanation quality EQ and general trust in the recommender systems GT have the same characteristics: their median and their upper quartile Q_3 is 4.0, that is, 50% of the scores are higher than or equal to 4.0, their lower quartile is 3.0, that is, 25% of the scores are lower than or equal to 3.0. Scores of 1 can be considered as outliers for EQ and GT if calculating lower outliers based on the interquartile range $\mathtt{IQR}$ as $Q_1 - 1.5 \times \mathtt{IQR}$. The perceived recommendation quality of the list RQL and the set RQS share the same value of 3.5 on average. The upper quartile Q_3 of RQL and RQS is 4.0. However, the lower quartile Q_1 of RQL is with 2.5 lower than 3.0 for RQS. Subsequently, a score of 1 can be considered as an inlier for the list but would be an outlier for the set if calculating lower outliers based on the interquartile range. Overall, the scores are higher and closer together for the set than for the list.

4.2 Evaluation of Subpopulations

In Section 3.3, we discussed demographic factors that could be associated with not following a study plan or a higher risk of dropping out. To evaluate the ratings, we performed the Mann-Whitney U test to identify any significant differences between complementary subpopulations (Table 3).

Table 3: Median scores by category (GT, EQ, RQL, RQS) by subpopulations (Aspect and Value). Mann-Whitney U test for the corresponding values of subpopulations: colored with 🟦 if $p <= 0.05$ and 🟧 if still statistically significant after correcting the p-values. Wilcoxon signed-rank test for the corresponding values of RQL and RQS: marked with * if $p <= 0.05$.

Aspect	Value	GT	EQ	RQL	RQS
Overall		4.00	4.00	3.50	3.50
P	**AR**	4.00	4.00	3.50	3.50
Program	**CM**	4.00	4.00	3.50	4.00
DD01	**2**	4.00	4.00	3.50	3.50
Semester	**n2**	4.00	4.00	3.50	3.75
DD02	**1 m**	4.00	4.00	3.50	3.50
Gender	**2 f**	4.00	4.00	3.50	4.00
DD03	**1 no**	4.00	4.00	3.50	3.50
Education BG	**2 yes**	4.00	4.00	3.00 *	3.50 *
DD04	**1 no**	4.00	4.00	3.50	4.00
Migration BG	**2 yes**	4.00	4.00	3.50	3.50
DD05	**1 no**	4.00	3.50	3.25	3.50
Study Plan	**2 yes**	4.00	4.00	3.50	3.50
DD06	**1 no**	4.00	4.00	3.00	3.50
Previous Credits	**2 yes**	4.00	3.50	3.50	4.00
DD07	**1 no**	4.00	4.00	4.00	4.00
Dropout Thoughts	**2 yes**	4.00	3.50	3.00 *	3.50 *

Across all subpopulations, we can observe that no value is below 3.0, that is, no group of students rates any category rather low. Investigating the subpopulations reveals that there are no differences in the median score regarding the general trust—it is 4.0 in all subpopulations—and that there are three slight differences regarding the explanation quality: in case of DD05, DD06, and DD07, the score is 3.5 in one group and 4.0 in the other. In terms of the recommendation quality of the list, the median scores of the subpopulations differ four times, and in terms of the set, six times. Both quality scores are in a similar proportion, that is, slightly higher in the same subpopulation, for DD06 and DD07. RQL has two further differences (DD03 and DD05), and RQS has four differences (study program P, DD01, DD02, DD04).

We found statistically significant differences in three cases, all concerning DD07, the question ”Have you ever thought about dropping out of your studies?” and the perception of the explanation quality EQ and the recommendation quality of the list RQL and the set RQS (highlighted by colors in Table 3). Students who have not thought about dropping out rated all three categories higher and these differences are significant, that is, not random. The difference for RQL remains significant even after the adjustment of all p-values using the Benjamini-Hochberg procedure (highlighted in orange in Table 3).

4.3 Evaluation of Variants

To compare the scores of the set and the list, we used the Wilcoxon signed-rank test and tested for a statistically significant difference between the median score of the recommendation quality of the list and the median score of the set. We first tested the ratings of all students before examining subpopulations (Table 3). Finally, we investigated the results in terms of a preferred variant that students selected in the survey.

Median Scores Overall and by Subpopulations. The median scores of the perceived recommendation quality of the list (RQL) and the set (RQS) reach an overall value of 3.5. Considering subpopulations, the median scores of the list and the set range from 3.0 to 4.0. We can observe that in 8 of 16 cases, RQL is equal to RQS (both are 3.5 or 4.0) and in the other 8 cases, the quality of the set is slightly higher than the quality of the list. In three subpopulations, RQL achieves a median score of 3.0: students whose parents or one parent have already studied (DD03 > 2 yes), students who have not taken credits for previous courses (DD06 > 1 no), and students who already thought about dropping out of their studies (DD07 > 2 yes)). A maximum score of 4.0 is achieved by only one subpopulation for RQL but by five subpopulations for RQS.

Comparing the ratings within subpopulations, the Wilcoxon signed-rank test indicates statistical significance (marked with * in Table 3) for students who are not first generation students (DD03 > 2 yes) and for students who have already considered to drop out (DD07 > 2 yes). None of these differences are still statistically significant after correcting the p-values.

Indirect Rating versus Direct Choice. Question 1 on our recommendation system (OS01) was about which variant the students thought was better for making a direct choice: 38 chose the list, 19 the set, 16 thought they were equally good, and 27 did not answer the question.

The box plots in Figure 2 show the distribution of median scores for the quality of recommendations of the list (RQL) and the set (RQS), based on the variant favored by the students. Students who would prefer the list (blue boxes) still rated the set not badly with a median of 3.5 while students who prefer the set rated the list with a worse median of 2.0 (yellow boxes). Students who rated the variants equally good (green boxes) indirectly rated the set better since the right

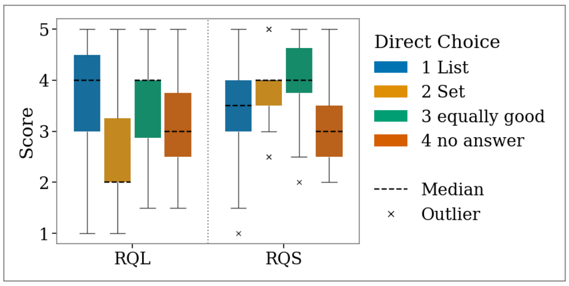

Figure 2: Distribution of the scores of RQL (boxes on the left) and RQS (boxes on the right) as box plots, colored by direct choice of a preferred variant (OS01).

green box is located higher than the left green box. Students who rated the variants equally good (green boxes) indirectly rated the set better than students who would prefer the set (yellow boxes) since the right green box is located higher than the right yellow box. Students who have not answered this question (orange boxes), indirectly rated RQL and RQS undecided with a median of 3.0.

We compared the ratings of subpopulations as in Section 4.2 and found statistically significant differences after adjusting the p-values in two cases: 1) for RQL and students who prefer the list and students who prefer the set, and 2) for RQS and students who prefer the set and students who have not answered OS01.

We compared the ratings of the list and the set and found a statistically significant difference after adjusting the p-values in the group of students who chose the set, but not in the other groups.

5 CONCLUSION

In this work, we present the results of a survey regarding a course recommender system aimed at supporting above all struggling students in their course enrollment for the second or third semester.

The results of the survey evaluation suggest that students tend to trust course recommender systems and that they tend to understand the simple explanations of how the recommendations are generated by the system presented in our previous work (Wagner et al., 2023) since the median for general trust (GT) and explanation quality (EQ) are 4 out of 5. These results are encouraging and promising.

Though students tend to understand well the recommendations presented in two variants, as a list and as a set, the median is in both cases 3.5, they rate the set presentation slightly better than the list presentation considering the distribution of the rating, see columns RQL and RQS in Table 3. Interestingly, the general trust in recommender systems and the understanding of the simple explanations are shared among all demographical subgroups. The analysis by subpopulations confirms the slight better rating of the set presentation; however, the differences are not statistically significant.

The answer to the question OS01 “Which variant do you think is better?” seems to give a different picture as 38 students prefer the list, 19 prefer the set and 16 think that both presentations are equally good. The more detailed analysis shows that the rating can be seen as contradictory: students who prefer the list still rate the set quite well with a median score of 3.5 and with no statistically significant difference while students who prefer the set rate the list rather poorly with a median score of 2.0 with a statistically significant difference. It is possible that the presence of a larger number of choices attracts more people (Bollen et al., 2010), as indicated by more students choosing the list. However, having such a large number of choices also increases the difficulty of making a decision, which can be reflected in the non-significantly lower rating of the set by the same students.

To summarize, the evaluation results are encouraging in terms of students’ overall trust in course recommender systems and their perception of the quality of the explanations. The study did not find a clear preference between presenting recommendations as a set or as a ranked list of courses. The evaluation in our previous work with historical data (Wagner et al., 2023) indicate that students at risk tend to enroll in more courses than the number of courses recommended to them. We interpret this finding as an advice for them to focus on less courses and pass them all. With this interpretation, recommending a specific number of courses as a set of courses would be more beneficial for students who are struggling, as opposed to providing a rank list.

Limitations. Since this was our first larger-scale survey of the recommender system, we compromised between the length of the survey and the number of items included. Although the number of valid questionnaires was not small, it was still not sufficient. Consequently, we were unable to thoroughly investigate the combinations of demographic factors, such as whether there was a statistically significant difference in the ratings of the list and set between students in the AR study program who had thoughts of dropping out and those who did not.

Future Works. Since the results of this survey with current students do not show a clear winner presenting the recommendations as a set or as a ranked list of courses, such a system could be implemented in two variants in future work: the list variant and the set variant. A/B testing could be performed to see if one system is preferred or is more successful. Future user surveys have the potential to delve into specific subpopulations that exhibited noteworthy results with statistical significance. Additionally, the contradictory results found comparing the indirect rating of the recommendation quality of list and set and the selected preferred variant should be investigated further.

ACKNOWLEDGEMENTS

During the preparation of this work, the authors used https://www.deepl.com and https://www.writefull.com in all sections to achieve better translations and more fluent texts. After using these services, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

REFERENCES

Berthold, C., Leichsenring, H., Brandenburg, U., Güttner, A., Kreft, A.-K., Morzick, B., Noe, S., Reumschüssel, E., Schmalreck, U., and Willert, M. (2012). CHE Diversity Report B1: Studierende mit Migrationshintergrund [CHE Diversity Report B1: Students with a migration background]. Technical report.

Bollen, D., Knijnenburg, B. P., Willemsen, M. C., and Graus, M. (2010). Understanding choice overload in recommender systems. In *Proceedings of the fourth ACM conference on Recommender systems*, RecSys '10, pages 63–70, New York, NY, USA. Association for Computing Machinery.

Elbadrawy, A. and Karypis, G. (2016). Domain-Aware Grade Prediction and Top-n Course Recommendation. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, pages 183–190, New York, NY, USA. Association for Computing Machinery.

Hernandez-Bocanegra, D. C. and Ziegler, J. (2023). Explaining Recommendations through Conversations: Dialog Model and the Effects of Interface Type and Degree of Interactivity. *ACM Transactions on Interactive Intelligent Systems*, 13(2):1–47.

Khan, M. A. Z. and Polyzou, A. (2023). Session-Based Course Recommendation Frameworks Using Deep Learning. In *Proceedings of the 16th International Conference on Educational Data Mining (EDM)*, Bengaluru, India. International Educational Data Mining Society.

Ma, B., Taniguchi, Y., and Konomi, S. (2020). Course Recommendation for University Environments. In *Proceedings of the 13th International Conference on Educational Data Mining (EDM)*, pages 460–466, Online. International Educational Data Mining Society.

Matayoshi, J. and Karumbaiah, S. (2021). Investigating the Validity of Methods Used to Adjust for Multiple Comparisons in Educational Data Mining. In *Proceedings of the 14th International Conference on Educational Data Mining*, pages 33–45, Online. International Educational Data Mining Society.

Miethe, I., Boysen, W., Grabowsky, S., and Kludt, R. (2014). *First Generation Students an deutschen Hochschulen: Selbstorganisation und Studiensituation am Beispiel der Initiative www.ArbeiterKind.de [First Generation Students at German universities: Self-organization and study situation using the example of the initiative www.ArbeiterKind.de]*. edition sigma.

Morsy, S. and Karypis, G. (2019). Will This Course Increase or Decrease Your GPA? Towards Grade-Aware Course Recommendation. *Journal of Educational Data Mining*, 11(2):20–46.

Pardos, Z. A., Fan, Z., and Jiang, W. (2019). Connectionist recommendation in the wild: on the utility and scrutability of neural networks for personalized course guidance. *User Modeling and User-Adapted Interaction*, 29(2):487–525.

Pardos, Z. A. and Jiang, W. (2020). Designing for serendipity in a university course recommendation system. In *Proceedings of the 10th International Conference on Learning Analytics & Knowledge (LAK)*, pages 350–359, New York, NY, USA. Association for Computing Machinery.

Polyzou, A., Nikolakopoulos, A. N., and Karypis, G. (2019). Scholars Walk: A Markov Chain Framework for Course Recommendation. In *Proceedings of the 12th International Conference on Educational Data Mining (EDM)*, pages 396–401, Montreal, Canada. International Educational Data Mining Society.

Statistisches Bundesamt (2023). Prüfungen an Hochschulen: Deutschland, Jahre, Nationalität, Geschlecht, Prüfungsergebnis [Examinations at universities: Germany, years, nationality, gender, examination result].

Urdaneta-Ponte, M. C., Mendez-Zorrilla, A., and Oleagordia-Ruiz, I. (2021). Recommendation Systems for Education: Systematic Review. *Electronics*, 10(14):1611.

Wagner, K., Merceron, A., Sauer, P., and Pinkwart, N. (2023). Can the Paths of Successful Students Help Other Students With Their Course Enrollments? In *Proceedings of the 16th International Conference on Educational Data Mining (EDM)*, pages 171–182, Bengaluru, India. International Educational Data Mining Society.