Yang, J. C. et al. (Eds.) (2018). Proceedings of the 26th International Conference on Computers in Education. Philippines: Asia-Pacific Society for Computers in Education

Personalized Guidance on How to Review Paper-based Assessments

Yancy Vance PAREDES[a*], I-Han HSIAO[a] & Yiling LIN[b]
[a]*Arizona State University, USA*
[b]*National Chengchi University, Taiwan*
*yvmparedes@asu.edu

Abstract: Providing feedback is one of the most effective methods to enhance student's learning. The absence of readily available data on paper-based assessments makes it impossible to know whether students received feedback and whether they have acted upon it or not. In our institution, we have been using a homegrown educational technology to support blended-instruction classes by integrating physical and cyberlearning analytics. Using the collected digital footprints of students, we were able to analyze their reviewing behavior. A study was conducted to investigate the effects of personalizing the reviewing sequence of paper-based assessments. Each student is presented with a personalized sequence of questions to review based on the importance of their mistakes. We found that the students who followed the suggested sequence improved significantly higher in the succeeding exam than those who reviewed the assessment arbitrarily or did not have any reviewing strategy. Results showed that personalized guidance on reviewing graded assessments effectively helped improve student performance.

Keywords: Personalized feedback, personalized learning, personalized guidance, programming learning, formal assessment, multimodal learning analytics.

1. Introduction

Paper-based assessment is a common tool to evaluate students' learning. It allows greater flexibility in preparing and administering the assessment. However, many desired and detailed learning analytics are unattainable. For instance, *how do students review the returned assessments*; *what are the impacts of the given feedback to their learning*, etc. While a range of efforts have been taken to support learning in online assessments (e.g. online submissions, auto-assessments, personalized feedback, etc.), little has been done to support personalized learning in blended learning environments, where classes utilize computer-assisted tools to support learning (e.g. online assignment submissions) and adopt paper-based formats for exams (e.g. to reinforce longhand behaviors (Mueller & Oppenheimer, 2014)). In our institution, we have been using WebPGA[1], a homegrown educational technology, to support blended-instruction class orchestration by integrating physical and cyberlearning analytics. Essentially, physical paper assessments are digitized, graded, and returned to students in an online environment. Using the collected digital footprints of students, we were able to assemble multimodal learning analytics and to analyze their reviewing behavior. In previous studies, we found that students exhibited diverse reviewing strategies. The success of good students was attributed to their determination to correct their mistakes (Hsiao, et al., 2017). Thus, in this paper, we hypothesized that students will benefit from a personalized reviewing sequence that is based on the urgency of the questions where they made mistakes. A classroom study was conducted to examine the pattern of differences of adaptive guidance in reviewing paper-based assessments.

The rest of the paper is organized as follows. Section 2 provides a discussion on the benefits of personalized guidance and adaptive feedback in learning. Section 3 illustrates an overview of the research platform and its design, as well as the data collection. Section 4 discusses our findings and the effect of using the personalized reviewing sequence to student performance. Finally, section 5 presents the conclusions and summarizes the work with future plans.

[1] https://cidsewpga.fulton.asu.edu/about/

2. Literature Review

2.1 Personalized Guidance in Learning

In the context of personalized learning, personalized guidance describes a group of techniques which provide a concise learning path to the learner. To implement personalized guidance in an intelligent educational system requires modelling the domain (learning content) and student interactions with the system (learning process). This enables the content to be presented or instructed in personalized sequences (Chen, 2008); or the learning process to be adapted to scaffold the learning activity (Azevedo & Jacobson, 2008). One of the common techniques in personalized guidance is Adaptive Hypermedia (AH), which utilizes the changes of the link appearances on the learning resources and guides students to the most appropriate ones (Brusilovsky, 1996). This approach relies on the synergy between the artificial intelligence (AI) of the system and the students' own intelligence, and often brings better results and higher satisfaction. Such technique has been evaluated and reported to help students to get to the right question at the right time, and significantly increase their chance to answer the question correctly in the self-assessment context (Brusilovsky & Sosnovsky, 2005; Hsiao, et al., 2010). The adaptive navigation support method has been further deployed in fusing social learning context. In an intelligent educational system with open social student modeling interfaces, greedy sequencing technique was adopted to maximize student's level of knowledge (Hosseini, et al., 2015a). The results revealed that the guidance increased the speed of learning for strong students, and improved the performance of students, both in the system and end-of-course assessments (Hosseini, et al., 2015b). Note that the mere presence of personalized guidance may not be sufficient to provide learning impact, what matters is whether the students choose to follow or to ignore the guidance (Hosseini, et al., 2015b).

2.2 Adaptive Feedback in Learning

Feedback is one of the most effective methods in enhancing student's learning (Hattie & Timperley, 2007). There is an abundance of factors that affect educational achievement. Some factors are more influential than others. For instance, feedback types and formats, timing of providing feedback, etc. (Shute & Zapata-Rivera, 2007). Studies have reported that positive feedback is not always positive for students' growth and achievement (Hattie & Timperley, 2007); "critical" rather than "confirmatory" feedback is the most beneficial for learning regardless of whether feedback was chosen or assigned (Cutumisu & Schwartz, 2016); content feedback achieves significantly better learning effects than progress feedback, where the former refers to the qualitative information about the domain content and its accuracy and the latter describes the quantitative assessment of the student's advancement through the material being covered (Jackson & Graesser, 2007). Several of the different feedback factors were explored on the intersections with the learner's variables (e.g. skills, affects) and reported to support personalized learning (Narciss, 2008). For instance, cognitive feedback was found to make a significant difference in the outcomes of both student learning gains in an intelligent dialogue tutor (Boyer, et al., 2008); student's affects were being adapted to improve motivational outcome (self-efficacy) (Boyer, et al., 2008; Dennis, et al., 2016); using student characteristics as tutoring feedback strategies to optimize students' learning in adaptive educational systems (Narciss, et al., 2014). While a large body of empirical studies investigate the impacts of feedback in the context of learning, we focused on researching adaptive feedback to guide students learn across physical and digital environment.

3. Methods

3.1 Research Platform

WebPGA was developed to connect the physical and the digital learning spaces in programming learning. This system facilitates the digitization, grading, and distribution of paper-based assessments. All actions performed by its users are logged. Examples of which include, but are not limited to: logging in and out, clicking an assessment to review, clicking a specific question to review, and using the

navigation buttons to move to another question. In certain cases, where it is applicable, the time spent performing the action is also recorded.

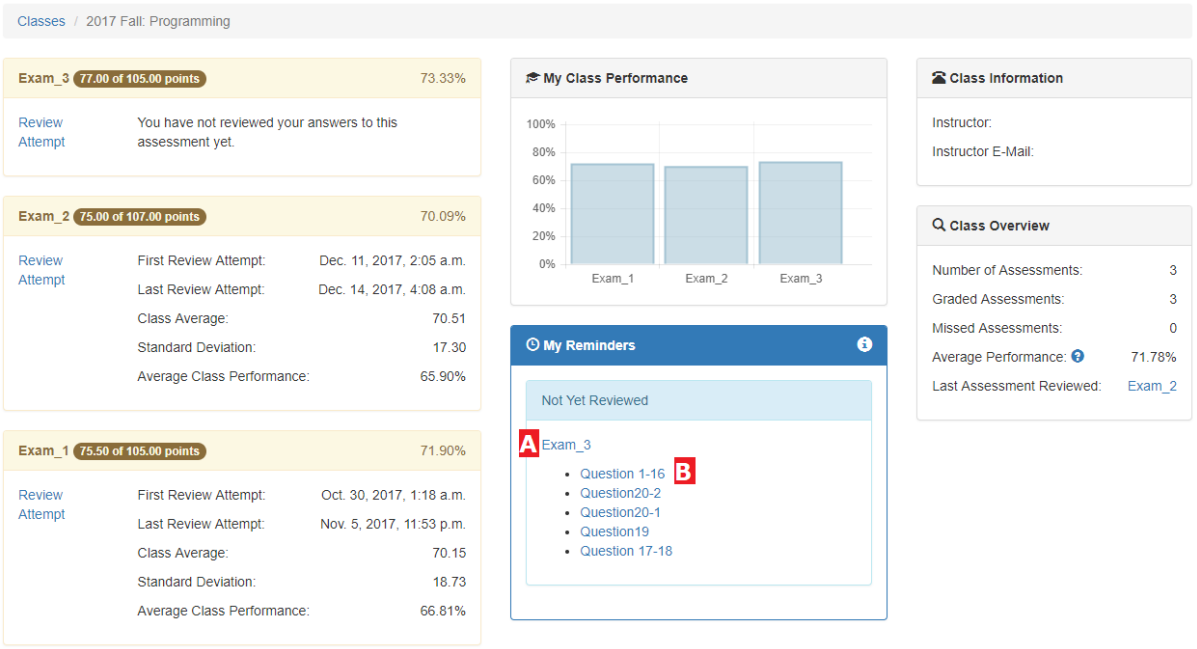

Figure 1. Student dashboard highlighting the reminders panel

Figure 1 presents the student dashboard which provides students an overview of their class performance. The leftmost panels list all the assessments along with the scores they obtained. It also informs them when they last reviewed an assessment, if applicable. Students can also click on a link to see detailed information about a particular assessment. The assessments are arranged from latest to oldest. In the center panel, a visualization is provided which gives students an overall picture of their progress in class. Below it is a newly implemented widget which provides students a personalized recommended reviewing sequence (further discussed in the next paragraph). The rightmost panels provide some administrative information about the class. Figure 2 presents an overview of an assessment which lists all the questions along with the scores obtained. A color coding scheme was utilized to make the presentation more meaningful. Green means the student got full credit, yellow means the student obtained partial credits, and red means the student did not obtain any credit. The questions are arranged based on how they were arranged in their physical paper counterpart. However, students can also opt to follow a recommended sequence provided by the system by clicking on the link on the upper right portion. Students can click on the thumbnail to review a particular question. This will open a panel (shown in Figure 3) where they can see more details about the question. This includes feedback from the grader, annotations on top of their scanned paper, detailed breakdown on how their answers were graded, and the score obtained for the question.

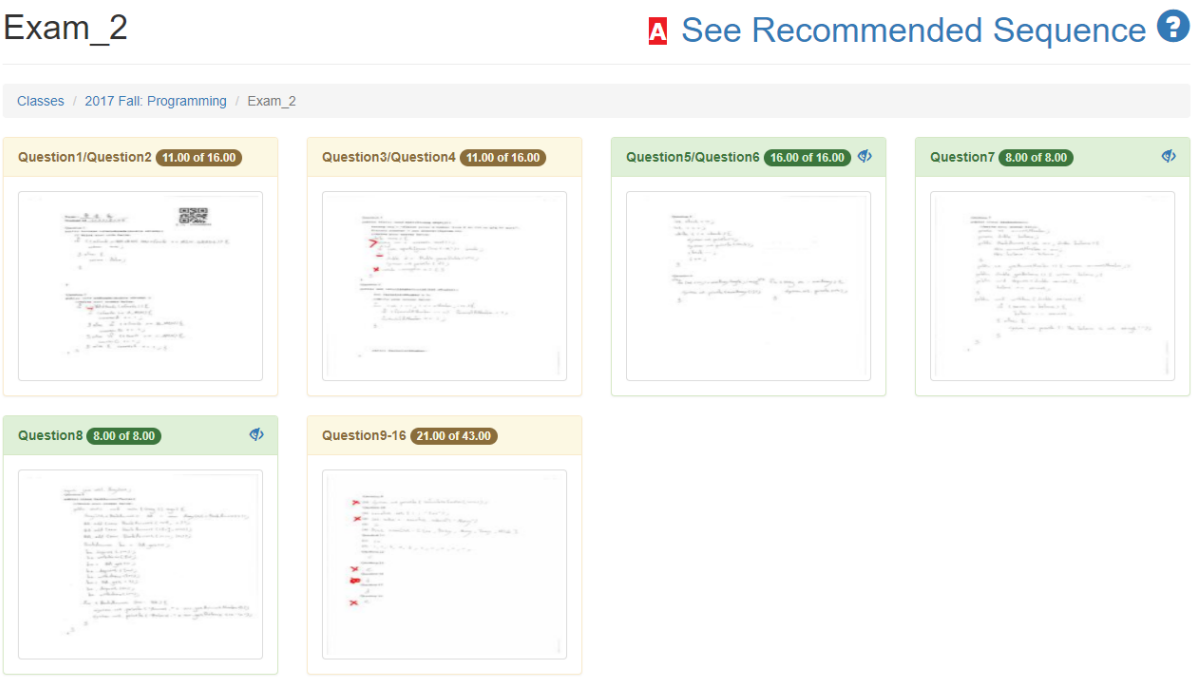

Figure 2. Assessment overview using the original sequence

Algorithm 1 Assessment and Question Listing in the Reminders Panel

```
1: procedure GetAssessmentsAndQuestions(S)
2:     for each A ∈ assessments of student S do
3:         Q := list of questions from A which are not yet reviewed
4:         for each q ∈ Q do
5:             q.normalized := q.raw_score_of_student / q.points_worth
6:             if q.normalized = 1 then
7:                 Remove q from Q
8:         if Q.isEmpty() then continue
9:         // Just in case there is a tie, use the next criteria
10:        Sort Q by q.normalized in ascending, q.points_worth in descending
11:    // Display the latest assessment on top
12:    Sort all assessments according to assessment_date in descending
```

Algorithm 2 Recommended Sequence

```
1: procedure GetRecommendedSequence(A)
2:     Q := questions from assessment A
3:     for each q ∈ Q do
4:         q.normalized := q.raw_score_of_student / q.points_worth
5:     // Just in case there is a tie, use the next criteria
6:     Sort Q by q.normalized in ascending, q.points_worth in descending
```

Personalization was introduced in the system, specifically on the student dashboard and in the assessment overview. Students are given personalized actionable reminders which list all assessments or questions that have not been reviewed (see Figure 1). The order of the items in the list is determined using Algorithm 1. If the student clicks on the name of an assessment from the list, they are redirected to the assessment overview (similar to Figure 2 but only different on how the questions are arranged) where questions are listed and arranged using Algorithm 2. On the other hand, if the student clicks on a specific question from the list, they are redirected to the question overview (see Figure 3).

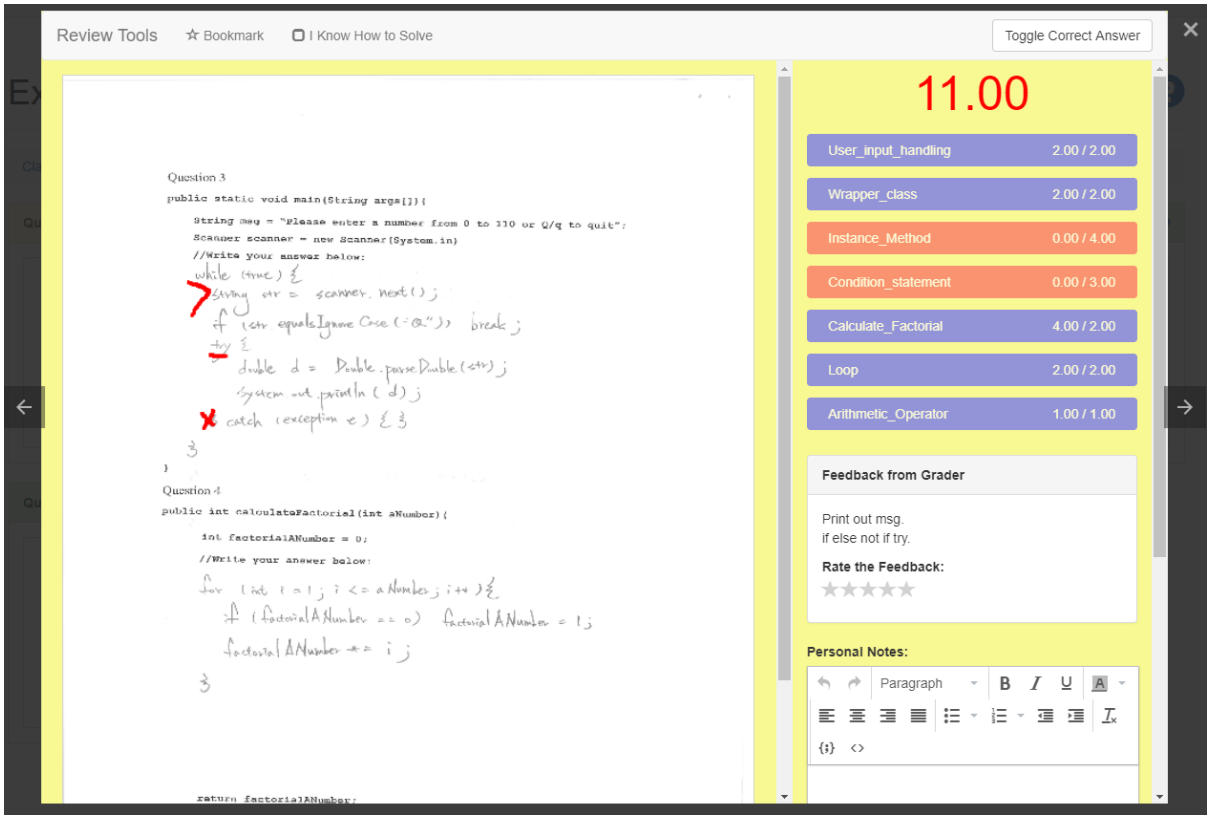

Figure 3. An overview of a single question which shows the scanned paper assessment (left panel) and the different feedback provided by the grader (right panel)

3.2 Data Collection

This study was conducted on an Introductory to Object-Oriented Programming course offered during the Fall 2017 semester. This course had 3 exams. Table 1 provides an overview of the students' performance in the exams. Among the 60 students enrolled, only 56 (93.33%) students were initially included in the study as those who dropped the course in the middle of the semester or did not take the three exams were excluded. All the students received the same instructions in class and used the same version of the system. After an exam was graded, it was released and made available to all the students at the same time. Students were given full autonomy to use or not use all the features of the system to help them review their graded exams.

Table 1
Overview of Class Performance

Exam	Student Count	Total Points	Average	Normalized	Std. Dev
Exam 1	59	105	70.15	66.81%	18.73%
Exam 2	57	107	70.51	65.90%	17.30%
Exam 3	57	105	77.11	73.43%	13.94%

The semester was divided into two equal time periods, namely *Exam1-Exam2* and *Exam2-Exam3*. Since personalization was only introduced to all students right after Exam 2 was administered in class, we focused this study on the *Exam2-Exam3* period only. A total of 1,959 user logs were captured by the system during this period. Afterwards, students were divided into two groups, namely *Guided* and *Not Guided*. If a student (1) clicked an assessment (Figure 1A) or a question (Figure 1B) from the list on the Reminders Panel, or (2) clicked on the "See Recommended Sequence" link on the Assessment Overview (Figure 2A), the student is classified under the *Guided* group. Otherwise, the student is classified under the *Not Guided* group.

Table 2
Improvement of the Two Groups during the Exam2-Exam3 Period

Group	Student Count	Average *Delta*	Std. Dev
Guided	24	0.09	0.66
Not Guided	16	0.03	0.78

The change in the normalized scores between Exam 2 and Exam 3, which will be referred to as *delta* (increase or decrease), was computed for each student. Students whose delta that are considered as outlier were excluded. The same was done for those who never logged into the system during the period. A total of 40 (66.67%) students were used in succeeding analyses. Table 2 summarizes the performance of the two groups during the period.

4. Results and Discussions

4.1 Learning Effects of Personalized Guidance

This study aims to integrate physical paper-based exams into a digital environment and provide personalized reviewing. Using the WebPGA platform, we were able to capture students' digital footprints which allows us to provide personalized guidance in maximizing student's learning. It is hypothesized that importance-based reviewing recommendation has a positive impact on learning. To verify this hypothesis, the *delta* between the two groups were compared. We found that, on average, the improvement of the students from the *Guided* group (9.23%) was significantly higher ($p<0.01$) than the students from the *Not Guided* group (3.18%). The Cohen's effect size is $d=0.08$. This finding prompted us to further investigate how the behavior of the two groups differ. We looked at two aspects of the groups, namely: their *skip distance* and their *reviewing effort*.

Table 3
Skip Distance of the Two Groups during the Exam2-Exam3 Period

Group	Student Count	Average *Skip Distance*	Std. Dev
Guided	24	2.06	0.99
Not Guided	16	1.03	0.77

4.1.1 Non-Sequential Reviewing Behavior

Whenever a student reviews an assessment, they are given the autonomy to choose a question to review regardless whether they followed the personalized recommended sequence or not. Students initially click a question to review from the assessment overview (Figure 2). This opens the question overview (Figure 3). Once they are finished, they could either (1) use the navigation buttons to go to the previous or the next question, or (2) close the current question and choose another question again from the assessment overview. When students use the personalized recommended sequence, important questions are identified and are presented to them. This results in non-sequential reviewing patterns. The number of questions in between the current question they are currently reviewing and the previous question they reviewed is referred to as *skip distance*. The order of the questions when it was administered in class was used as reference. We found that when students followed the recommended sequence, they had a significantly higher ($p<0.01$, $d=1.1$) skip distance than those who did not. The non-guided group obtained an average skip distance of 1.03, which is not necessarily a bad thing (an average skip distance of 1 indicates that the student simply reviewed the questions of the assessment in a sequential manner). This might also indicate the inability of the students to identify which questions to skip so that they can focus on questions that needs to be reviewed right away. Table 3 summarizes the average skip distance of the two groups. Lastly, we found that there is a weak positive correlation ($r=0.06$) between the average skip distance of a student and his or her improvement. However, it is not statistically significant. This result aligned with the literature that process feedback (suggested review sequence) can only do so much. What matters is the content feedback (concept correctness feedback) (Jackson &

Graesser, 2007). Thus, we further inspected what were the content that the students reviewed and the subsequent effects in Section 4.2.

4.1.2 Reviewing Effort

In addition to students' reviewing behaviors, we also looked into their reviewing efforts. Through the help of personalization, students become aware of the importance of certain questions based on their performance. We hypothesized that this will enable students to become more persistent in reviewing those questions to get them right the next time. This means that they will review certain questions multiple times. Also, this will prompt them to attend to these questions immediately and therefore be able to cover more mistakes when reviewing. To verify these hypotheses, we counted the number of times each student reviewed questions during the entire period. Also, for each student, we counted the number of questions from Exam 2 that the system would identify as important. Then, among these questions, we counted how many were reviewed by the student. This ratio will be referred to as the student's *review coverage*. There was no significant difference between the number of times the students from the *Guided* group (x=12.21, s=11.69) and those from the *Not Guided* group (x=8.56, s=6.40) reviewed. When the review coverage by those from the *Guided* group (x=79.44%, s=31.56%) and those from the *Not Guided* group (x=67.50%, s=47.26%) were compared, no significant difference was found as well. Although no significance was found in reviewing efforts, it must be noted that the personalization was only introduced after the first exam. It will be interesting to know whether the same trend can be seen if multiple assessments, such as quizzes, are also available.

4.2 Improvement on Knowledge Components

To further understand in finer details the effect of the personalization on the performance of the students, we focused on the students from the *Guided* group and inspected all the knowledge components associated to all the questions of Exam 2 and Exam 3. Only those that are common in the two exams were considered. Among the 50 knowledge components, 10 were excluded from the analysis. Afterwards, the knowledge components were grouped based on Java Ontology as defined by Hsiao et al. (2010). A total of 5 topics were identified. These are listed below along with the number of knowledge components classified under them.

T01. Basic Programming Concepts (8)
T02. Object & Class Concepts (20)
T03. Control Structure (3)
T04. Iteration (7)
T05. I/O Handling (2)

For each topic, the raw scores were added up then divided by the maximum possible points. This is done for both Exam 2 and Exam 3. Table 4 summarizes the average performance of the students in the topic for the exam. A topic followed by an asterisk (*) indicates that it is statistically significant ($p<0.05$). Among the 5 topics, 3 were the topics where the students who used the personalized review sequence had a significant improvement. Looking at the general trend, students did better on topics which were already covered in the past.

Table 4
Effects of Personalization on Common Topics

Exam		**T01***	**T02***	**T03**	**T04**	**T05***
Exam 2	Average	0.56	0.74	0.78	0.69	0.46
	Std. Dev	0.14	0.13	0.15	0.15	0.39
Exam 3	Average	0.89	0.80	0.75	0.71	0.85
	Std. Dev	0.08	0.12	0.29	0.39	0.28
Delta		0.33	0.06	-0.03	0.02	0.40
Cohen's d		2.9	0.5	0.1	0.1	1.1

Finally, we wanted to know how the system helped different types of students. Students from the *Guided* group were split into two. The median score of Exam 2 (x=74) of the entire class was used as

the cut-off point to classify the students as either *Stronger student* (11 students) or *Weaker student* (13 students). Table 5 provides an overview on how the two groups improved during the period which is also visualized in Figure 4. We found that for fundamental topics, such as T01, weaker students had a significantly higher improvement of 3.72%. The same can be seen in a more complex topic, such as T05 at 5.38%. Also, both have large effect sizes. This finding is interesting as it shows that through personalized and adaptive guidance, weaker students can take full advantage of focusing on reviewing their "weakness" and consolidate their fundamental knowledge, which could subsequently lead to the improvement on both simple and advanced topics.

Table 5
Improvement on Common Topics of the Guided Group

Exam	T01*	T02	T03	T04	T05*
Stronger student	0.28	0.05	-0.09	0.05	0.23
Weaker student	0.37	0.08	0.02	0.00	0.54
Cohen's d	1.1	0.2	0.3	0.1	0.9

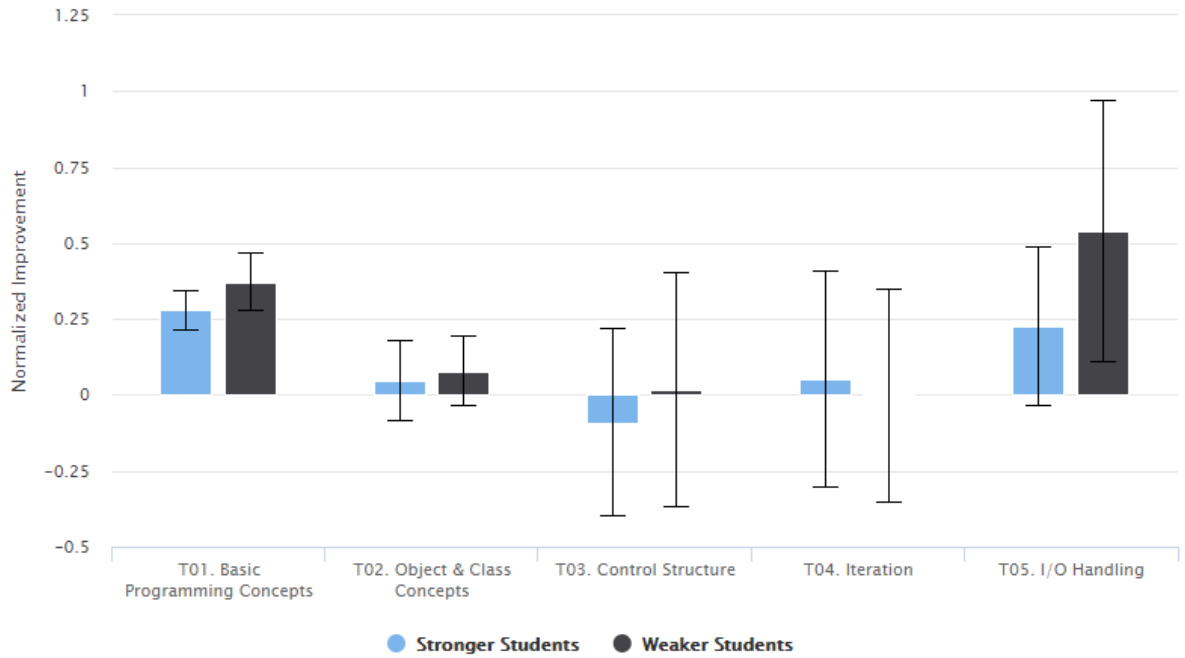

Figure 4. Improvement according to topics within the *Guided* group

5. Conclusion

5.1 Summary

Using the research platform developed, empirical data on how students review graded paper-based assessments were collected and analyzed. A study was conducted on an Object-Oriented Programming course. Students were grouped into two based on whether they used the personalized review sequence recommended by the system during the *Exam2-Exam3* time period. Students who used it had a significant higher improvement in their succeeding exam compared to those who did not. We investigated the behaviors of the two groups in terms of their reviewing patterns and their reviewing effort. We found that those who used the personalized reviewing sequence had a higher skip distance, which indicates that the system was able to provide them a personalized sequence which informs them what questions are important. However, in terms of students' reviewing efforts, although no significance was found, a common trend was seen for both the number of times they reviewed a question and their review coverage. Students were able to review questions multiple times and were able to review questions that were identified as important. Knowledge components which were

common to the two exams during the time period were identified and grouped together into topics. We found that students who followed the personalized reviewing order improved significantly. Additionally, we took a closer look on the magnitude of the improvement. Overall, students were found to improve on a range of topics. The findings suggest that the personalization effectively guides both strong and weak students to attend to the right questions when reviewing. The effect was more apparent in weaker students, as they improved more in fundamental topics (T01) as well as more complex ones (T05). Based on the findings in this study, the implemented personalization looks promising in helping students improve their review strategy.

5.2 Limitation and Future Work

Since the personalized recommended review order was only introduced in the middle of the semester, only logs from the second half of the semester were analyzed. Moreover, only the data from two exams were looked into. As a result, we did not find any significance in personalization effects in reviewing efforts. In the future, more exhaustive analysis should be done, particularly the progression of assessments (e.g. all the quizzes in between exams) in Computer Science Education (CSE) courses.

References

Azevedo, R., & Jacobson, M. J. (2008). Advances in scaffolding learning with hypertext and hypermedia: A summary and critical analysis. *Educational Technology Research and Development, 56*, 93-100.

Boyer, K. E., Phillips, R., Wallis, M., Vouk, M., & Lester, J. (2008). Balancing cognitive and motivational scaffolding in tutorial dialogue. *International conference on intelligent tutoring systems*, (pp. 239-249).

Brusilovsky, P. (1996). Methods and techniques of adaptive hypermedia. *User modeling and user-adapted interaction, 6*, 87-129.

Brusilovsky, P., & Sosnovsky, S. (2005). Individualized exercises for self-assessment of programming knowledge: An evaluation of QuizPACK. *Journal on Educational Resources in Computing (JERIC), 5*, 6.

Chen, C.-M. (2008). Intelligent web-based learning system with personalized learning path guidance. *Computers & Education, 51*, 787-814.

Cutumisu, M., & Schwartz, D. L. (2016). Choosing versus Receiving Feedback: The Impact of Feedback Valence on Learning in an Assessment Game. *EDM*, (pp. 341-346).

Dennis, M., Masthoff, J., & Mellish, C. (2016). Adapting progress feedback and emotional support to learner personality. *International Journal of Artificial Intelligence in Education, 26*, 877-931.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of educational research, 77*, 81-112.

Hosseini, R., Hsiao, I.-H., Guerra, J., & Brusilovsky, P. (2015a). Off the Beaten Path: The Impact of Adaptive Content Sequencing on Student Navigation in an Open Social Student Modeling Interface. In C. Conati, N. Heffernan, A. Mitrovic, & M. F. Verdejo (Ed.), *Artificial Intelligence in Education* (pp. 624-628). Cham: Springer International Publishing.

Hosseini, R., Hsiao, I.-H., Guerra, J., & Brusilovsky, P. (2015b). What Should I Do Next? Adaptive Sequencing in the Context of Open Social Student Modeling. In G. Conole, T. Klobučar, C. Rensing, J. Konert, & E. Lavoué (Ed.), *Design for Teaching and Learning in a Networked World* (pp. 155-168). Cham: Springer International Publishing.

Hsiao, I. H., Huang, P. K., & Murphy, H. (2017). Integrating Programming Learning Analytics Across Physical and Digital Space. *IEEE Transactions on Emerging Topics in Computing, PP*, 1-1. doi:10.1109/TETC.2017.2701201

Hsiao, I.-H., Sosnovsky, S., & Brusilovsky, P. (2010). Guiding students to the right questions: adaptive navigation support in an E-Learning system for Java programming. *Journal of Computer Assisted Learning, 26*, 270-283.

Jackson, G. T., & Graesser, A. C. (2007). Content matters: An investigation of feedback categories within an ITS. *Frontiers in Artificial Intelligence and Applications, 158*, 127.

Mueller, P. A., & Oppenheimer, D. M. (2014). The pen is mightier than the keyboard: Advantages of longhand over laptop note taking. *Psychological science, 25*, 1159-1168.

Narciss, S. (2008). Feedback strategies for interactive learning tasks. *Handbook of research on educational communications and technology, 3*, 125-144.

Narciss, S., Sosnovsky, S., Schnaubert, L., Andrès, E., Eichelmann, A., Goguadze, G., & Melis, E. (2014). Exploring feedback and student characteristics relevant for personalizing feedback strategies. *Computers & Education, 71*, 56-76. doi:https://doi.org/10.1016/j.compedu.2013.09.011.

Shute, V. J., & Zapata-Rivera, D. (2007). Adaptive technologies. *ETS Research Report Series, 2007*.