# Web-Based Programming Grading Assistant:

# An Investigation of the Role

# of Students Reviewing Behavior

by

Po-Kai Huang

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved May 2017 by the
Graduate Supervisory Committee:

I-Han Hsiao, Chair
Brian Nelson
Kurt VanLehn

ARIZONA STATE UNIVERSITY

August 2017

ABSTRACT

Paper assessment remains to be an essential formal assessment method in today's classes. However, it is difficult to track student learning behavior on physical papers. This thesis presents a new educational technologyWeb Programming Grading Assistant (WPGA). WPGA not only serves as a grading system but also a feedback delivery tool that connects paper-based assessments to digital space. I designed a classroom study and collected data from ASU computer science classes. I tracked and modeled students' reviewing and reflecting behaviors based on the use of WPGA. I analyzed students' reviewing efforts, in terms of frequency, timing, and the associations with their academic performances. Results showed that students put extra emphasis in reviewing prior to the exams and the efforts demonstrated the desire to review formal assessments regardless of if they were graded for academic performance or for attendance. In addition, all students paid more attention on reviewing quizzes and exams toward the end of semester.

# DEDICATION

Dedicated to my beloved parents, Wen-Chung and Meei-Fang who offer unconditional love and support. Thank for giving me the best education you could.

ACKNOWLEDGMENTS

I take this opportunity to thank my thesis advisor Dr. Sharon Hsiao. She helped me in need whenever I got stunk in research. She allowed this paper to be my own work, but guided me the direction whenever I needed it. I am grateful to Dr. Brian Nelson and Dr. Kurt VanLehn for lending their time to be member of my committee.

I am also thankful to all CSI Lab members and my friends who helped and guided me in my thesis study. I am especially appreciated Yancy Paredes who assisted me in research and writing.

I would have never today's achievement without my father- the wise father and loving mentor in my life. His support and faith inspired me every decision I made. All my accomplishments dedicate to you. Thank you Dad.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

# Chapter 1

# INTRODUCTION

We have begun to see more and more educational technologies emerging nowadays (i.e. smart classrooms etc.). Examples of which include Clickers (19) and multi-touch tabletops (13), etc. Even with all of these new technologies, most data sources of students performance are collected from computer-assisted formative assessments or retrieved from learning management systems. Less is focused on integrating multi-modal learning analytics, from physical to digital activities. In this thesis, the goal is to design and study a new educational technology that will bridge physical and cyber learning spaces. Moreover, this thesis aims to study the impacts of the technology on students learning.

## 1.1 Motivation

In today's blended learning environment, paper-based exams are still one of the most popular formal assessment methods. Paper exam provides the teacher a reasonable high degree of flexibility in making them with any text editing software. On the contrary, online assessments may require instructors to learn new content authoring tools, which may not only limit the choices of the software but also lead to a higher time cost in learning to use them. Additionally, paper-based exams may reduce the potential for academic dishonesty compared to its online counterpart. However, Paper-based exams have drawbacks.For instance, as the class size increases, grading becomes more challenging. There are usually many inconsistencies in the grading (among and within the graders) (8); there are difficulties in providing feedback (handwritten feedback is time consuming; delivering graded paper exams back to students

can be challenging etc.). Therefore, graders end up providing only limited feedback on tests; as a result, students usually end up focusing mostly on their final scores, among several other issues (1). From the literature, we have learned that feedback is one of the most effective methods to enhance students learning (7). Finally, it is difficult to perform learning analytics because of the absence of data. With the above mentioned reasons, a new educational technology was designed that would harness the benefits of traditional learning activities (i.e. paper exams), and would enable the performance of advanced digital learning analytics.

## 1.2 Research Questions

In this work, a new educational technology is designed to facilitate grading paper-based assessment items, providing feedback and delivering graded results to students via an online platform. I hypothesize that providing a digital channel, which allows students to access their physical assessments, will have a chance to promote review and reflection, and positive impact on learning. Thus, the main focus of this work is to investigate students reviewing behaviors and to find out *how these behaviors make progress toward learning* Furthermore, this work aims to answer the following specific research questions:

- *Do students care about their returned exams at all?*
- *Will students be able to learn to be more accountable in monitoring their progress?*
- *When they do or they dont, what are the behavior changes and the impacts?*
- *How do they adjust their strategies?*

- *What are the learning effects when students focus on the summative feedback (i.e. final scores) as opposed to the formative feedback or the detailed graded items?*

In the rest of the thesis, the content is structured in the following chapters. In Chapter 2, the theoretical background supporting the educational technology design is presented. Chapter 3 describes the research platformWPGA in detail. Chapter 4 discusses the methodology, classroom study design, and data collection and evaluation measures. Chapter 5 presents evaluation results and discusses implications. Finally, Chapter 6 summarizes contribution, conclusions and future work of this research.

# Chapter 2

# LITERATURE REVIEW

This section discusses related literature, which includes feedback on learning, technology support in feedback. It also discusses results from previous analysis and similar technique. In addition, it tells about some previous data analysis result direct or redirect to lead the study in following chapter. This thesis is extended analytics of two papers: *The Role of Reviewing Formal Assessments in Programming Learning* (15) and *Uncovering Reviewing and Reflecting Behaviors From Paper-based Formal Assessment* (10). Therefore many of content and methods are derived from those papers.

## 2.1 Role of Feedback in Learning

There are plenty of factors which affect academic achievement. This includes learning experience, learning feedback, teaching style and etc. Each of which may also influence one another. Many of these factors are not easy to quantify. Several papers highlight the importance of feedback in learning. In their paper, Hattie and Timperley (7) discuss what constitutes an effective feedback, which may not necessarily be positive or negative. They also found out that positive feedback does not always have positive impact to students' academic performance. The same is true for negative feedback. For instance, compliments like *Excellent!* , *Good explanation* are ineffective feedbacks; negative feedbacks like *unclear on method description* , *not enough explanation on* which are more constructive feedbacks to students. In educational data mining paper, Cutumisu (4) also suggests that the quality of feedback is most beneficial no matter what feedback results come out.

In paper (7), researchers also stressed on importance of timing feedback. Another studyPeerStudio (12) also has same focus: fast feedback is very important. The fast feedback is easy to catch student attention because of easy to reflect their difficulties on current learning. The article also mentions that the influence to students on slow feedback and no feedback is no significant difference. Moreover, studies also show that availability of immediate self-corrective feedback increase efficiency on reviewing examinations (5); Students are much more beneficial from feedback which scores on individual components of an assignment than from feedback with summed up scores. Students can learn their misunderstanding part directly (12). Overall, above studies advocate the importance of timing feedback.

In recent learning analytics literature, I found that student's assessment grades can be source of predictor which can predict their whether can end up in good academic performance or bad. Moreover, aggregated data sources are keys to getting timely and predictive feedback (18). Therefore, in my thesis, the goal is to streamline feedback delivery into digital world, capture students' performance on their learning feedback, and understand how those feedback impact on student' learning behaviors.

## 2.2 Technology Support in Feedback Generation and Delivery

Automated assessment is one of the most popular methods in scaling feedback generation. Methods guarantee fast feedback to deliver students. Idea like programming Integrated Development Environment can give user direct feedback after they compile their codes. Such techniques have already been widely used in many educational fields, such as programming, mathematics, physics and etc., to build up assessments for students. Exemplar systems are WEB-CAT (6) and ASSYST (11), among many. The common approach is to apply pattern-matching techniques that verify students' answers by comparing them with the correct answers. Unfortunately, in our domain of

interest, programming learning, automatic programming evaluation emphasizes only the solely aspects of an answer. It cannot provide personalized answers. For example the application doesn't judge student's answer in the way whether the answer logic is correct or not. Instructor need to spend extra time to review those questions after automated assessment. Therefore, it is still a challenge for machine to judge logical/knowledgeable questions. Under this concern, the paper-based examination has its own benefits. Here comes our research question: *how could we integrate feedback across space?* One example is the tablet grading system (2). It uses tablet scanners to digitalize the paper assessment and provide grade interface to assist grading works on table. It introduces benefits of digitizing paper exams: default feedback can be kept in digital pages; hide students' information from graders to prevent potential bias. Overall, the field of automatic evaluation is less focused on grading paper-based programming problems. Our goal is to study students' learning effectiveness through the use of feedback delivery tools.

## 2.3 Data Model Analytics in Architecture

Learning data is broad and complex. It takes time to train data in desire modeling format. Different research direction led the different form of data. The paper (10) submitted to Learning Analytics & 2017 Knowledge Conference describes the scenario of high performing students have a different review strategy to low performing ones. The sequential data modeling  Hidden Markov Model (HMM) is method applied to see patterns. The HMM generate a serial states of reviewing and reflection behavior of students.

The result of HMM is in Figure 2.1, the left side is high performing student learning process; right side is low performing one. The both learning process look very similar (both group review behavior start with reviewing exam) but different

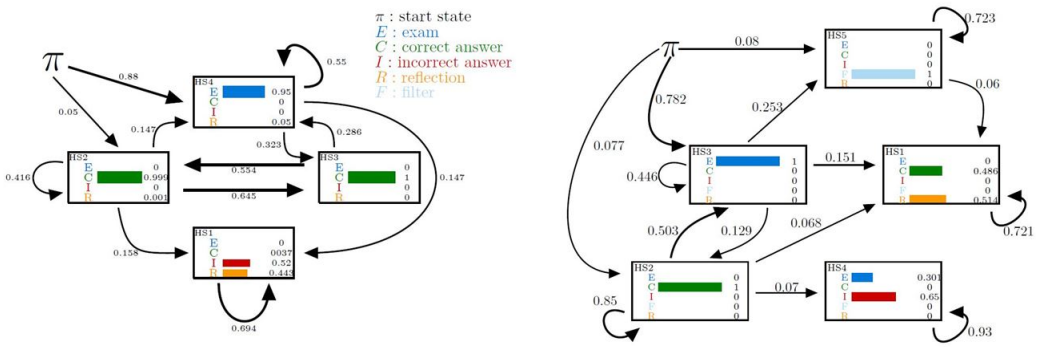


**Figure 2.1:** Hidden Markov Model for High Level Student and Low Level Student Respectively

in reflection strategy. The high performing students show that they will reflection immediate when they review incorrect questions. On the other hand, low performing ones also review incorrect questions but they fail to do reflection for them. This is critical difference to distinguish two groups. I found this pattern from data of first exam in data structure course. However, the data has limitation and bias: the system just launches and has some flaws inside; data logs consist of one exam. Therefore, the pattern is not obvious to see when modeling data of whole semester. It encourages several evaluations on behavioral change, performance analytics in following chapters.

## 2.4 Difference to Similar Platform: Gradescope

The education technology is not the latest fashion. The different technology applies for different purpose. However, if the general goal is similar, it is inevitable that two education applications have similar features. Gradescope (17) is the application which helps to instructor on grading paper-based assessment. The designs and grading methods are similar to WPGAdigitizing paper assessment, grading exams with the Web application. Gradescope has been in use for four years, which serves mature grading interfaces and visual analytics to aid instructors or graders on grading exams or assignments. Gradescope emphasizes user experience of instructors or

graders and try to reduce the loading of grading as low as possible. Similarly, WPGA presents comparable interfaces to decrease grading work, but it is just being one of many benefits of WPGA. Instead of focusing on instructors or grading, WPGA is designed for students. The main focus is to help elevate student academic performance. Specifically, in this thesis, the analysis emphasis is on student review behaviors and strategies to their learning.

## 2.5 Idea of Programming Grading Assistant

The idea of blending physical world with digital space is not novel. The system I will introduce in following chapters originate from Programming Grading Assistant (PGA) (14). PGA is mobile education application which uses QR-codes and OCR (8)(Optical character recognition) to recognize student information and hand-writing. Instructors and graders expect to grade on mobile devices which is convenient without heavy loading on carrying paper exam and benefit in storage students examination. The paper points out on grading consistency between different graders and it has user study on contrast of grading with PGA and tradition paper grading. The major improvement from tradition paper grading is that the grading consistency is better in PGA than tradition way. However, the application face great challenge with OCR technology, it requires optimal lighting condition for recognition and also whole procedure from digitized data to graph is time-consuming the cost to have recognizable digital data is too high to popularize. Those limitations inspire of development of web-based programming grading assistant.

# Chapter 3

# RESEARCH PLATFORM: WEB PROGRAMMING GRADING ASSISTANT (WPGA)

A web-based system was designed to facilitate the grading of paper-based exams and the delivering of feedback online. The name of the system is Web Programming Grading Assistant (WPGA). WPGA connects physical paper-based assessments into the digital space which ensures teachers the flexibility to continue using paper exams without having to learn new content authoring tools. WPGA features several functionalities for both instructors and students, three main key elements are: (1) Documenting paper-based assessments; (2) Augmented grading and feedback-giving; (3) Reflective feedback delivery.

## 3.1 Traditional Paper Grading vs WPGA in Feedback Delivery:

Before telling into WPGA system design, the idea inspired us developed WPGA educational system is researches on traditional paper grading and feedback delivery. Figure 3.1 illustrates the flow of events when grading paper-based exams in a traditional setup. In the past, instructors produced paper examinations and collected them back after administering the exam. Instructors are often faced with bottleneck during the grading process since it cannot be done in parallel. Furthermore, communication with the graders present a high cost as well. These results to a slow delivery in feedback. This raised an interesting question: *how could the time before feedback could be given to the students be decreased?* More specific, *how to reduce cost from paper collection and communication?* The WPGA system attempts to address these.

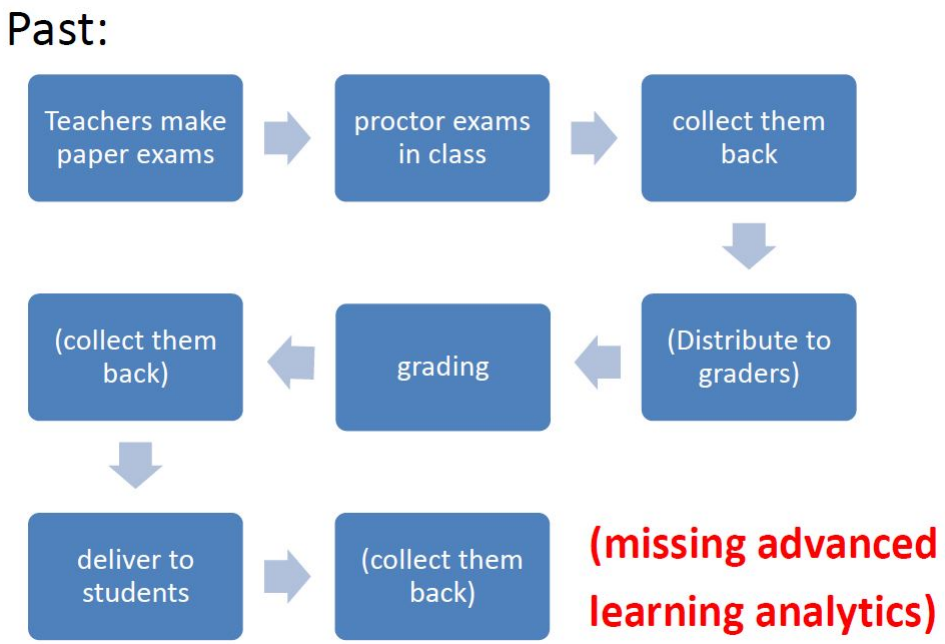


**Figure 3.1:** The Process of Traditional Paper-Based Exam Feedback Delivery

Figure 3.2 illustrates the flow of events when grading paper-based exams using WPGA. The amount of time before feedback is given to students is shorter. Instructors could assign questions from an exam to graders for them to grade. These reduce time on paper collection and communication which was used to tell to graders face-to-face. There is no need for physical access to the papers because the graders do the grading in the system– improving the their efficiency. Also, students view their scores and feedback in the system. This setup eliminates the possibility of academic dishonesty among students where they attempt to make modifications in their graded paper. Finally, in the traditional setup, it is difficult to do learning analytics, hard to track student learning process from all in paper format since there is no way to capture how students review their graded exams.

The WPGA system provides a digital student interface which captures the behavior of the students which enables the performing of advanced learning analytics which will be introduced in following chapters.

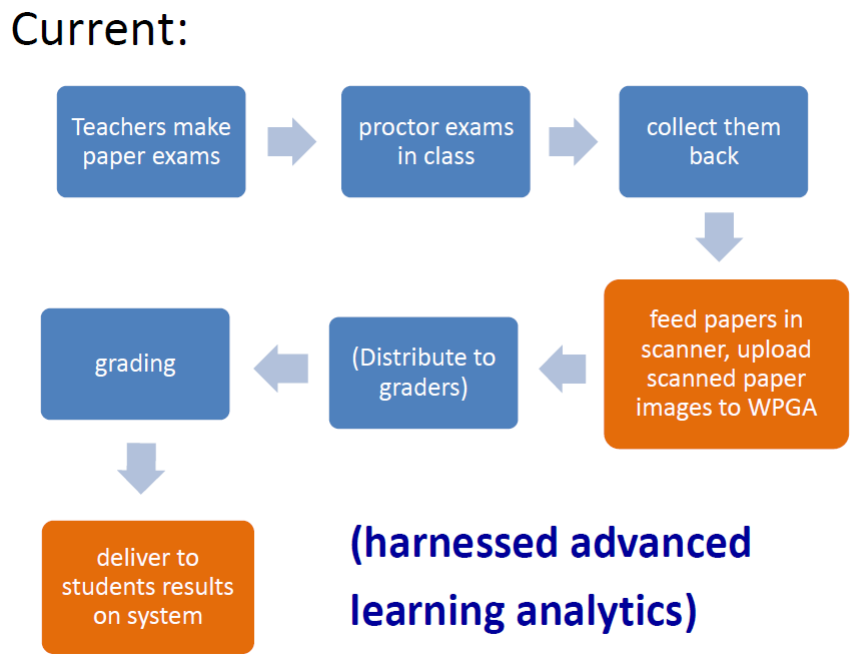


**Figure 3.2:** The Process of WPGA Feedback Delivery

## 3.2 Digitalizing Process of Paper-Based Assessments

Quick response codes (QR-codes) were utilized to label and identify a hard copy exam of a student. Instructors will use document feeder to scan all students' paper exams. Afterwards, all the scanned exams will be uploaded to WPGA database and stored as images. The digitizing process not only transforms physical content into digital version, but also establishes a link between student and their page exam.

## 3.3 Augmented Grading Interfaces

In the grading interface, there are two levelsmanagement level and grading level. From the instructor's point of view, the students' paper exam are labeled during

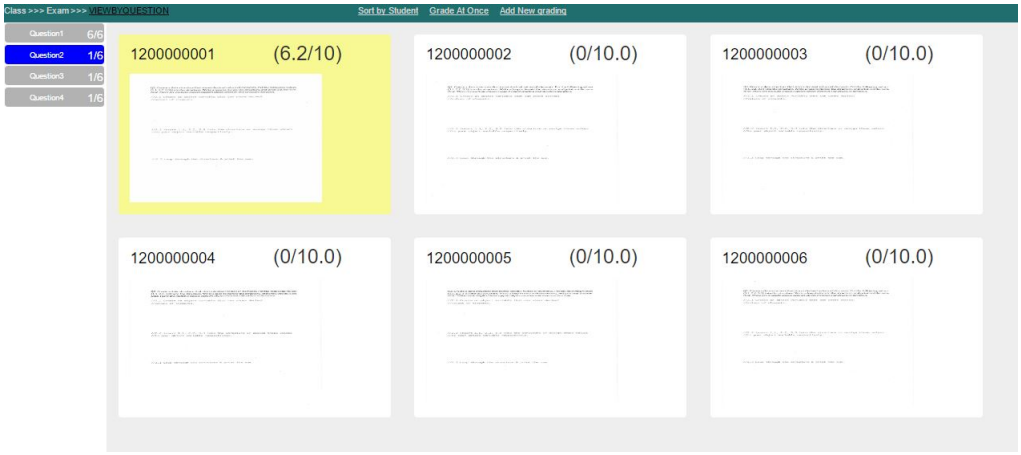


**Figure 3.3:** Exam Level View in Grading Interface

digitizing process (question number, question text, exam number etc.). The process makes it easy for the instructor to assign a question to be graded by a particular grader. In effect, multiple graders are able to grade different questions of the same exam simultaneously. Also, this improves consistency in grading.

The grading interface provides features that ensures efficient and consistent grading. There are two levels of view: Exam level and Question level. Graders are able to focus on grading only the questions that are assigned to them. From the grader's aspect, the sorting feature in exam level allows a grader to easily grade the same question on different students' exams all at once (Figure 3.3). This resolves the challenge of having to flip through hundreds of pages when grading stacks of paper exams and also increases grading coherence because each grader focuses on grading a certain question (8).In addition, there is special feature for grading a quiz if it is recorded only for attendance *Grade At Once*. It automatically gives full credits to all the papers of the students.

The question level view is where the grading takes place (Figure 3.4). The biggest feature is the interactive feedback buttons (on the upper right corner). The buttons are associated to learning contents, along with the grading scheme. Each question

defaults to a perfect score, with all feedback buttons being blue (full understanding). For each click, the grades are automatically recalculated based on instructor's preconfigured grading schemes. The feedback button will turn red (partial understanding) or grey (missed this concept). The idea is that a question is associated to a specific course topic. The graders grade a question based student's understandings of those topics.Additionally, graders can type feedback comments in the text area.

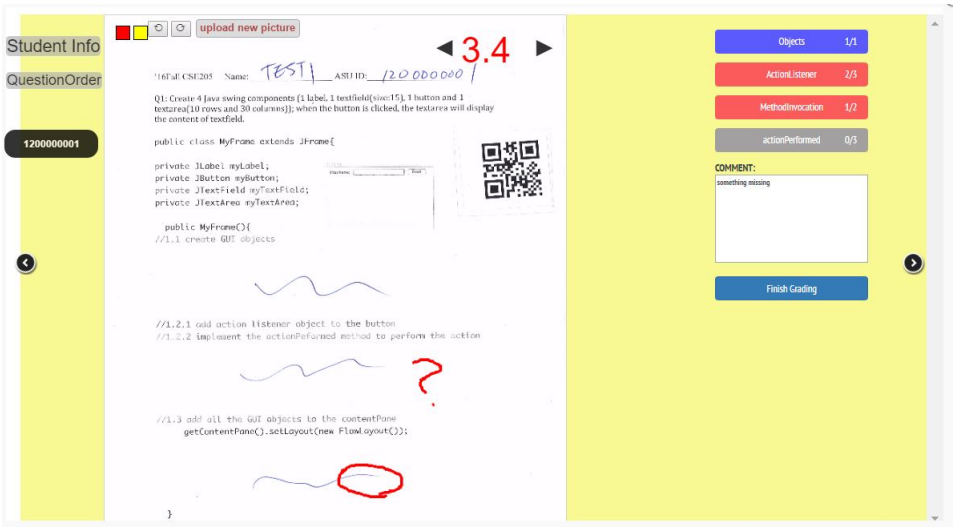


**Figure 3.4:** Question Level View in Grading Interface

According to studies previously conducted (8; 9), graders prefer to type in comments rather than physically writing them on paper because of convenience to reuse the same comment to similar errors. Having to reuse earlier comments is another benefits using WPGA for grading. The grading interface has more features and more complex than student one. Therefore I design a help page for instructors and graders to learning how to grading with WPGA (Figure 3.5).

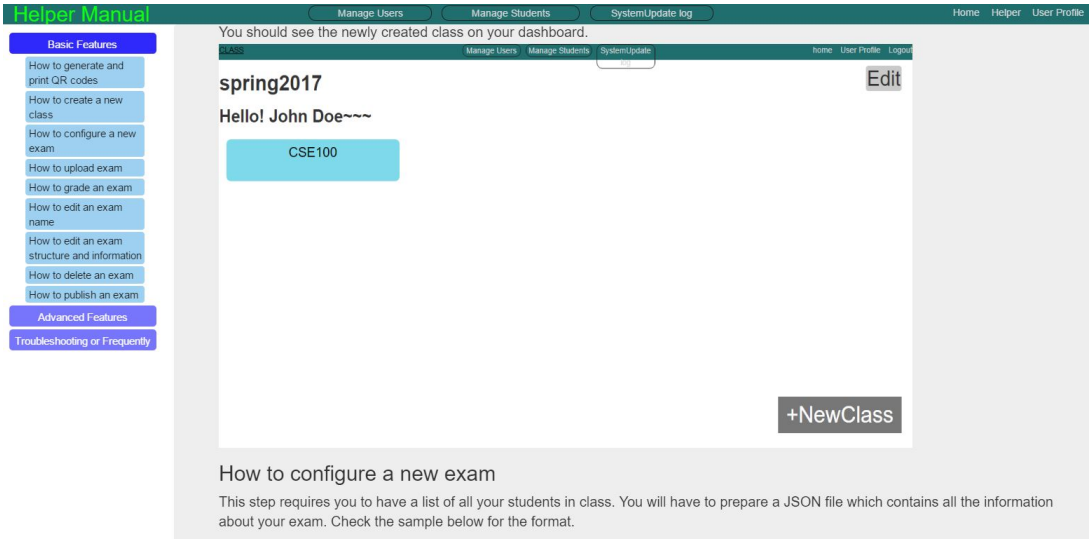


**Figure 3.5:** The Help Page for Instructors and Graders

## 3.4 Reflective Feedback Delivery: Student Interfaces

To directly benefit students on learning and achieve learning analytics, a student interface for feedback delivery is essential. The interface provides students basic review feature, such as viewing their exam scores, and reviewing graded questions. It also has advanced features available: self-reflection and monitoring of their performances via note-taking, bookmarking, and explicitly acknowledging their understanding (Figure 3.7). There are three forms of reflection prompts: (a) a star bookmark to note the importance of or the need to reference a question in the future; (b) a checkbox to express I know how to solve it now to indicate questions that the student have learned; and (c) a free form text area where the student can type elaborated notes. Such features to reflect (3) can encourage students to do self-learning on their responses, and self-reflect on the reasoning processes that led to a deep learning experience. The collection of bookmarks, checkboxes, and notes are considered as the source of what student learned, and thus he/she might become more metacognitively aware of his/her own subject matter knowledge (16).

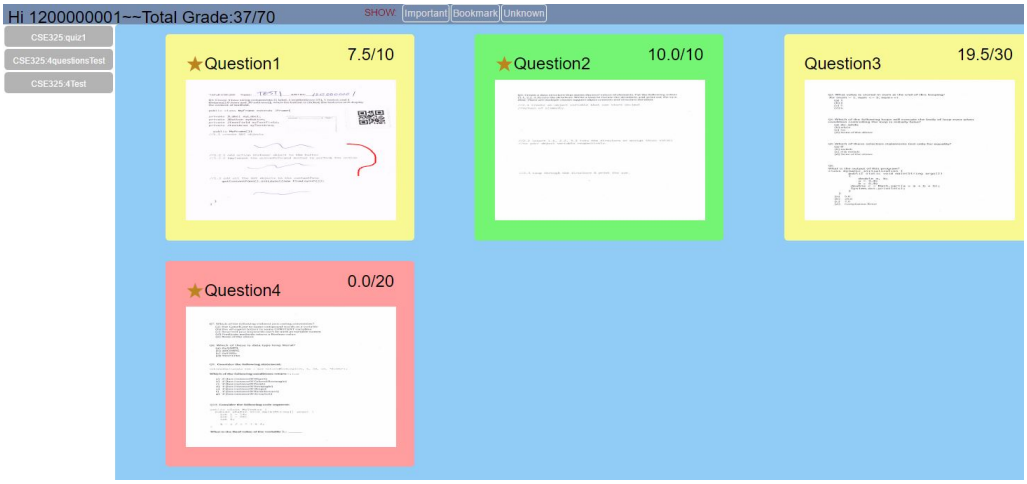


**Figure 3.6:** Exam Level Overview in Student Interface

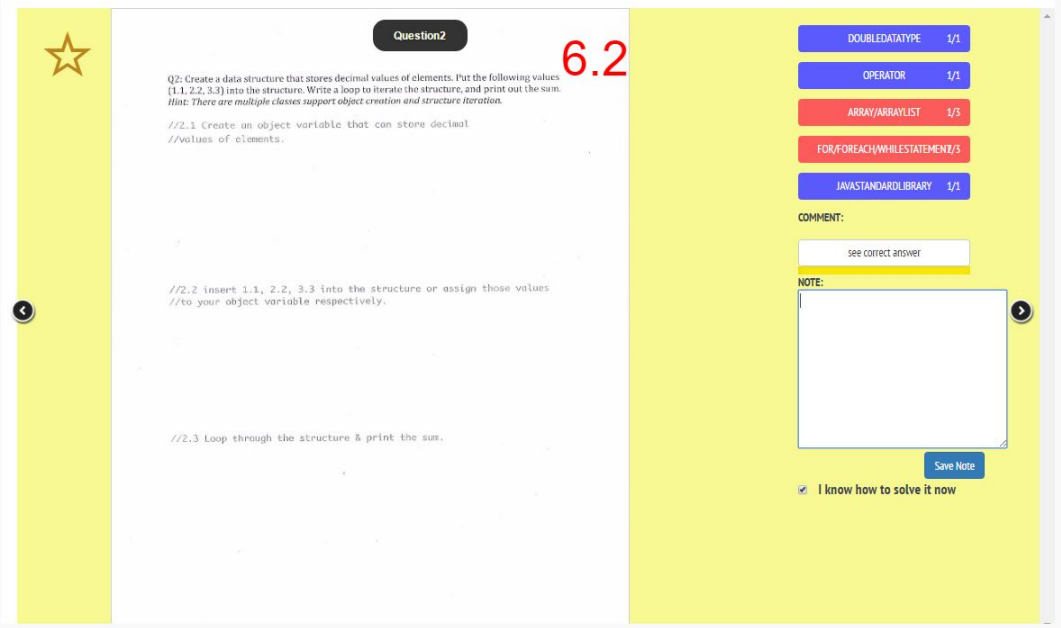


**Figure 3.7:** Question Level View in Student Interface

There are two levels of views in the student interface: Exam level view (Figure 3.6) & Question level view (Figure 3.7).They are similar to the grading interface but have different functionalities. Exam level view provides the overview result on quiz or exam, which displays the overall grade scores (shown on upper left corner), bookmark (star mark on screenshot of question) and the number of incorrect/correct questions. The colors of questions facilitate navigation: green shows full marks, red indicates zero marks, and yellow indicates partial credit. Additionally, student can filter questions by bookmark, question unknown or both by clicking button on upper middle of interface. To review on specific questions, one can click on the question snapshot and enter the Question level view. In this level, one can see more details on the question: the score obtained, the grading scheme, the grading feedback, and the correct solution. In addition, students can take notes to reflect on the particular problem-solving assessment item or also bookmark the question for future reference. A checkbox that students click indicates whether they already know how to solve the problem after reviewing it. This is particularly useful for questions where they committed mistakes (Figure 3.7 bottom right corner). The system logs all the actions of the students. These actions include the logging in the system; viewing the feedback given by the grader; reviewing a question. Eventually, I use student logs to do analysis and return result as feedback to aid students on their learning. Above information and operation also be described in help page of student interface (Figure 3.8).

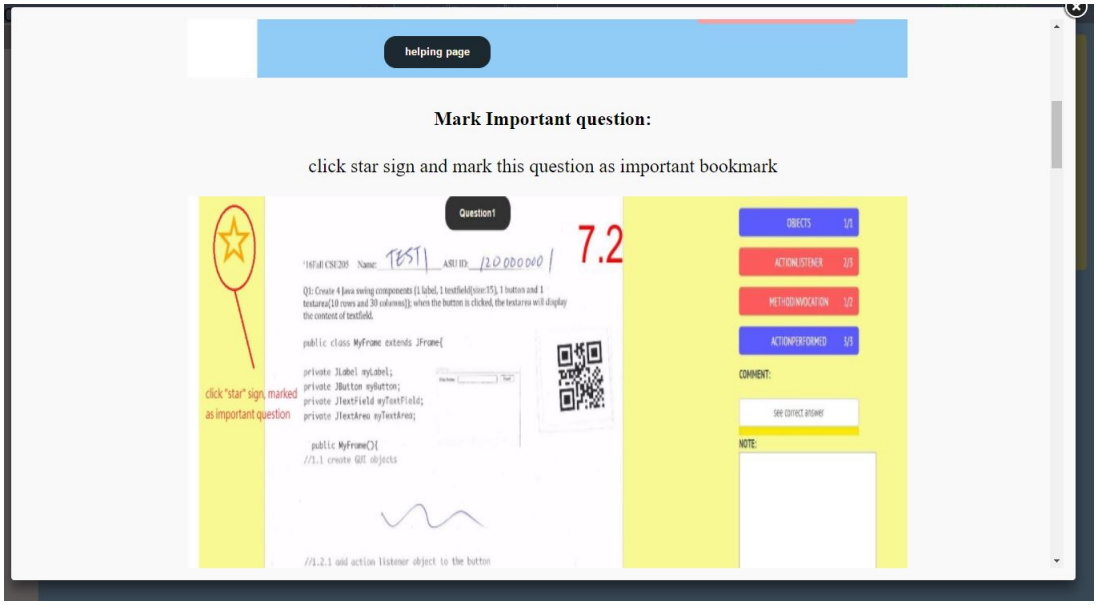


**Figure 3.8:** Helping Page in Student Interface

# Chapter 4

# METHODOLOGY

In this section, the methods used to understand the reviewing behavior of students are discussed. Based on the preliminary study results (10), HMM models revealed that there was a difference in the reviewing behaviors of the A & B students. Therefore, this thesis continues delving in students reviewing behaviors, specifically, on examining students behavior changes.

## 4.1 Study Design:

A classroom study was conducted in an undergraduate level course– Data Structure and Algorithms (CSE 310) in 2016 Fall semester. To investigate students monitoring and reviewing behaviors and learning effectiveness, focus was given on their performance and grades changes.

### *4.1.1 Data Collection*

There were a total of 33,738 action logs made by 247 students. The course was taught in traditional blended instruction format face-to-face course, online submitted assignment, and in-class paper-based quizzes including three exams and thirteen quizzes. Among all thirteen quizzes, six of them were for credit while the remaining were recorded only for attendance. Also I organized data with different labeling system which will be discussed in 4.2 section. To see student behavioral change, the whole semester split into two time periods based on three exams– the first exam to the second exam and the second exam to third exam, they are denoted as *Exam1-Exam2* and *Exam2-Exam3* respectively. There were five reviewing and reflecting behavioral

| Behavior | Action | Description |
|---|---|---|
| Review | Exam | Click on Exam tab to examine each individual quiz/exam; Overall marks are shown. |
| | Correct Question | Click on a single question to examine question & answer details; Question marks, grader's / instructor's feedback; Question is color coded in green |
| | Incorrect Question | Click on a single question to examine question & answer details; Question marks, grader's /instructor's feedback, and reflection prompts are shown; Question is color coded in yellow or red |
| | Filter | Click on any advanced filters to select targeted set of questions, i.e. show only bookmarked questions, not yet reflected questions, show both. |
| Reflect | Reflect | Keep notes on the question reviewing interfaces; Bookmark the question for future review; Tick a checkbox to acknowledge one's understanding on a question. |

**Table 4.1:** WPGA Behavioral Actions Descriptions

actions logged in WPGA: review exam, review correct question, review incorrect question, filter and reflect. The detail action descriptions are summarized in Table 4.1. In this thesis, only the direct reviewing activities were considered, which include all the actions of reviewing correct and incorrect questions.

### *4.1.2 Descriptive Data*

WPGA was launched during the Fall 2016 semester. At that time, six computing courses were using it. This includes Introduction to Programming, and Data Structure and Algorithms. There were 35 active users registered in the system (3 instructors and 32 student graders). WPGA was able to collect more than 90,000 logs throughout the semester.

In this study, only students who participated in all three exams were considered. Out of the 247 students, only 239 of them used WPGA at least once. These eight students were then omitted from the analysis. During the Exam1-Exam2 time period, a total of 210 students used the system. However, only 189 of them reviewed Exam1 prior to taking Exam2. During the Exam2-Exam3 time period, a total of 210 students used the system. There were 198 of them who reviewed either Exam1 or Exam2 prior to taking Exam3.

## 4.2 Data Labeling

The students academic performances, as well as their behavior changes, were obtained to investigate the impacts of the students reviewing behavior on their learning.

### *4.2.1 Performance Labels*

The average of the three exams was used to represent the overall academic performance of a student. The median of the class academic performance (X=81.67) was used as the threshold to classify a student as either High-Level or Low-Level groups of students.

### *4.2.2 Behavioral Change Labels*

To analyze behavioral pattern differences and learning, I applied different labels to represent behavioral changes across *Exam1-Exam2* and *Exam2-Exam3* time periods. According to the exam score ranges, students were labeled as A, B, and C, where A represents 90 and above, B represents 80 to 90, and C denotes to 80 and below. Based on these letter grades and exam periods, students are further classified into *Improving, Retaining,* and *Dropping* groups. For an instance, a student who scored a letter grade B in the first exam, and had improving grades to letter A in the second exam will be labeled as *Improving12*; one who scored a letter grade B in the first exam and remained the same in the second exam will be labeled as *Retaining12*; finally whoever scored a letter grade B in the first exam and dropped to letter grade C in the second one will labeled as *Dropping12*.

### *4.2.3 Review Action Labels*

There were a total of N=5,907 review actions. Additionally, to distinguish the actions from reviewing or just skimming the questions, the median of the duration of all review actions (X=14 seconds) was used as the threshold to classify a review action either as Deep Review or Shallow Review. Therefore, there are four review actions *Deep Review* or *Shallow Review*. Therefore, there are four review actions *deep correct question review, deep incorrect question review, shallow correct question review, shallow incorrect question review.*

## 4.3 Quantifying Student Behavioral Change

Each review action is timestamped. The amount of time spent and the time distribution of review attempts were used to measure the behavioral change.

### *4.3.1 Observation of Review-Time Chart*

A review-time chart was used to visualize the reviewing behavior distribution of the students (Figure 4.1). The x-axis in the chart represents time (in days) in a given period. On the other hand, the y-axis represents a student. The students were sorted in descending order according to their average final scores. Each blue point corresponds to a particular review action. The chart for the two time periods (Exam1-Exam2 and Exam2-Exam3) was plotted separately. Furthermore, the chart for the two student groups (Improving and Dropping) was plotted separately as well.

It can be observed that students who belong to the Improving group performed more review actions compared to those who belong to the Dropping group. In addition, some students, mostly from the Improving group, did review consistently which led to the formation of the horizontal solid lines. In order to validate these observations, a deeper review pattern analysis is performed.

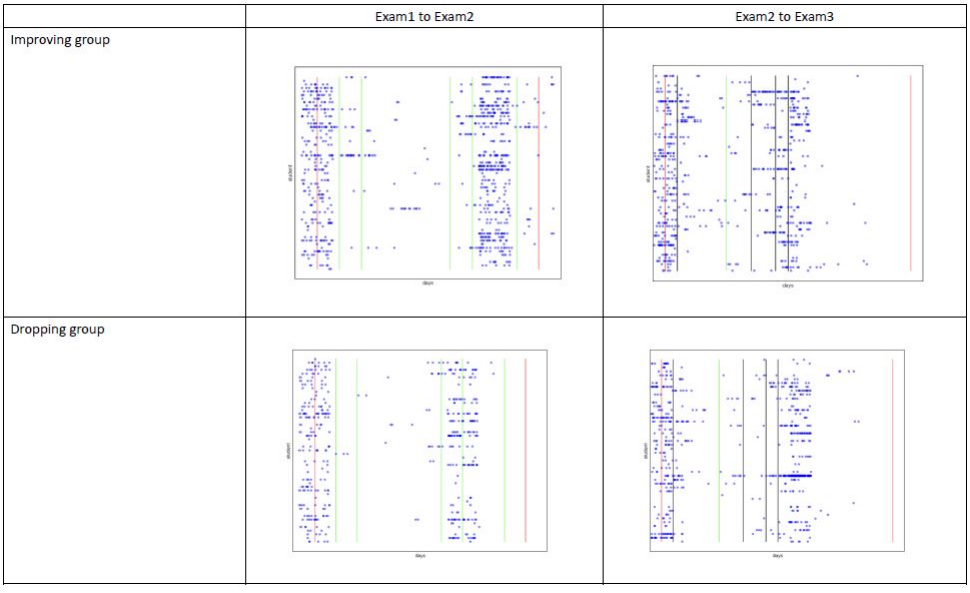


**Figure 4.1:** The Review-Time Chart

### *4.3.2 Review Pattern Modeling*

To see how student attentive on their graded exams or quizzes, the first review attempt was considered. Table 4.2 shows the exam review coverage (which refers to the number of students who reviewed an exam or quiz) and the average first attempt review time. From table 4.2, students appeared to gradually attempt to review sooner toward the end of semester which inspired the question to see how student vigilant on review their exams and quizzes.

| Exam | Number of Students Who Reviewed | % of Students Who Reviewed | Time Before First Review |
|---|---|---|---|
| Exam1 | 219 | 91.63% | 4.6 Days |
| Exam2 | 215 | 89.96% | 2 Days |
| Exam3 | 190 | 79.5% | 0.8 Day |

**Table 4.2:** First Review Attempt Table for Exams

To see the pattern, the First Review Attempt Table (Table 4.3) had been plotted. The table shows the elapsed time (in seconds) between the time an exam or quiz was published by the instructor and the first review attempt of the students. Each row of the table corresponds to a student while each column corresponds to an assessment (exam or quiz). It can be observed that several students did not review some exams or quizzes(blank cells in Table 4.3). To see how many exams and quizzes had been reviewed, coverage analysis had been employed. The review coverage of each student was computed using the Formula (4.1) to determine the usage rate of WPGA. It was found out that half of the exams and the quizzes were reviewed at least once by at least one student (M=0.56, SD=0.27).

| user_id | quiz1_ | quiz2_ | quiz3_ | exam1_ | quiz4_ | quiz5_ | quiz6_ | quiz7_ | quiz8_ | exam2_ | quiz9_ | quiz10_ | quiz11_ | quiz12_ | quiz13_ | exam3_ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 664 | | 1412695 | | 96965 | | | | | | 163141 | | | | | | 447843 |
| 661 | | | | 1217456 | | | | | | | | | | 419899 | | 1092 |
| 495 | | 3656854 | | 5583547 | | | | | | 2565415 | | | | | | 666 |
| 696 | 1959603 | 881112 | 253155 | 2008 | 2720705 | 5075188 | 3875365 | 2076646 | 2171866 | 24160 | 948076 | 572873 | 436212 | 236997 | 327294 | 1005 |
| 698 | | | | 2054 | | | | | | 22386 | | 696051 | | | | 1929 |
| 674 | | 3892016 | | 2544 | | 750102 | 844829 | 432544 | 290643 | 17830 | | | | | | 849 |
| 577 | | | | 1972 | | | | | 363228 | 23982 | | | | | | |

**Table 4.3:** First Attempted Review Time Table (Uncolored Value is Blank Cell)

$$Review\ Coverage = \frac{number\ of\ exam\ and\ quiz\ reviewed}{total\ number\ of\ exams\ and\ quizzes} \tag{4.1}$$

Due to above explanations, data imputation on first review attempt table had been performed. Ideally, those missing values (or blank cell) should be filled in by positive infinity. However, for this analysis, the missing values were instead filled in by using the maximum value of that particular student plus a certain fixed value (X=1000) as shown in Figure 4.4 (uncolored cell). Moreover, all blank cells setting same maximum value is unfair. It could let two student time list be much similar each other because of same value which may cause bias. The aim of analyzing this table is to know how students put effort on reviewing their exams and quizzes.

| user_id | quiz1_ | quiz2_ | quiz3_ | exam1_ | quiz4_ | quiz5_ | quiz6_ | quiz7_ | quiz8_ | exam2_ | quiz9_ | quiz10_ | quiz11_ | quiz12_ | quiz13_ | exam3_ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 664 | 1413695 | 1412695 | 1413695 | 96965 | 1413695 | 1413695 | 1413695 | 1413695 | 1413695 | 163141 | 1413695 | 1413695 | 1413695 | 1413695 | 1413695 | 447843 |
| 661 | 1218456 | 1218456 | 1218456 | 1217456 | 1218456 | 1218456 | 1218456 | 1218456 | 1218456 | 1218456 | 1218456 | 1218456 | 1218456 | 419899 | 1218456 | 1092 |
| 495 | 5584547 | 3656854 | 5584547 | 5583547 | 5584547 | 5584547 | 5584547 | 5584547 | 5584547 | 2565415 | 5584547 | 5584547 | 5584547 | 5584547 | 5584547 | 666 |
| 696 | 1959603 | 881112 | 253155 | 2008 | 2720705 | 5075188 | 3875365 | 2076646 | 2171866 | 24160 | 948076 | 572873 | 436212 | 236997 | 327294 | 1005 |
| 698 | 697051 | 697051 | 697051 | 2054 | 697051 | 697051 | 697051 | 697051 | 697051 | 22386 | 697051 | 696051 | 697051 | 697051 | 697051 | 1929 |
| 674 | 3893016 | 3892016 | 3893016 | 2544 | 3893016 | 750102 | 844829 | 432544 | 290643 | 17830 | 3893016 | 3893016 | 3893016 | 3893016 | 3893016 | 849 |
| 577 | 364228 | 364228 | 364228 | 1972 | 364228 | 364228 | 364228 | 364228 | 363228 | 23982 | 364228 | 364228 | 364228 | 364228 | 364228 | 364228 |

**Table 4.4:** First Review Attempt Table after Data Imputation

### *4.3.3 Kullback-Leiber Divergence*

The Kullback-Leiber divergence measures the divergence between two non-symmetric distributions. The formula is presented in Formula (4.2): two distributions Q and P, the equation indicates the divergence from distribution Q to P. A discrete distribution is defined as a students first attempted review across all assessments overtime. There-

fore, the divergence denotes to the relative entropy from one student to another. In this work, Kullback-Leilber divergence matrix was computed according to the first-attempt-to-review based on the behavior changes (improving and dropping groups) and the different time periods, table 4.5. Each row is a list of degree of Kullback-Leilber divergence from the distribution of one student to other students distribution. Higher degree of Kullback-Leiber divergence suggests more diverged from one student to another.

$$D_{KL}(P \parallel Q) = \sum_i P(i)\ log\frac{P(i)}{Q(i)} \tag{4.2}$$

| | | | | |
|---|---|---|---|---|
| 0.2776 | 0.277839 | 0.074567 | 0.011182 | 0.472747 |
| 0 | 2.86E-08 | 0.242789 | 0.178164 | 0.64021 |
| 2.86E-08 | 0 | 0.242896 | 0.178271 | 0.640317 |
| 1.095825 | 1.096477 | 0 | 0.061286 | 0.524318 |
| 0.788782 | 0.789301 | 0.061928 | 0 | 0.461 |
| 1.225128 | 1.225639 | 0.456978 | 0.450817 | 0 |
| 0.687864 | 0.688329 | 0.062171 | 0.000136 | 0.460824 |
| 0.317131 | 0.317392 | 0.0715 | 0.008274 | 0.469695 |
| 0.280096 | 0.280012 | 0.562215 | 0.615488 | 0.639181 |

**Table 4.5:** The Part of Kullback-Leilber Matrix

# Chapter 5

# EVALUATION

This section presents a series of evaluation results including effectiveness evaluation, efficiency evaluation, and the subjective evaluation. The goal is to answer the core research questions: *how much did students engage in reviewing? how efficient would the student be able to review? how does the review strategy affect student's academic performance?*. Effectiveness evaluation examines different learning strategy on the pattern differences and use Analysis of Variance (ANOVA) to see difference among behavioral label groups. Efficiency evaluation with Kullback-Leiber Divergence is evaluated to see student attendance of review and efficiency on reviewing with learning curve. Subjective evaluation shows student learning experience with WPGA in whole semester.

## 5.1 Student's Efforts in Reviewing

It was found out that throughout the entire semester, the students from the High-Level group of students (M=20.11, SD=23.26) had significantly ($p < 0.01$) fewer review actions, on average, compared to those from the Low-Level group (M=35.58, SD=35.23). Results were initially counterintuitive because it is expected for High-Level students to do more review to ensure in high performance status. However, it seems that higher amounts of review actions does not necessarily mean that they were reviewing significantly more effectively. A possible explanation could be that WPGA records every action, which assuming a student clicks on a question, s/he reviews it. However, student could either skim it after the click or click on a question intended to review, but doing other thing instead. High-Level students may have lesser review

actions but might have spent more time reviewing each question. Additionally, since they received higher grades, they should have fewer incorrect questions to review, and subsequently have fewer review actions. Therefore, to inspect the level of engagement in reviewing and how effective of the review actions impact on learning, I further looked into efficiency of review strategies and impacts of behavior changes.

## 5.2 Influence of Behavior Change in Reviewing

The amount of time spent by students in reviewing correct and incorrect questions were investigated. In this analysis, focus is only given to the Deep Correct and Deep Incorrect reviews. Since this analysis focuses mainly on behavior changes, only the students from the Improving and the Dropping groups were considered. The evaluation has been submitted to *the thirteenth International Computing Education Research Conference* , which is currently under review.

### *5.2.1 Improving Group Effectively Reviewed*

The normalized amount of time spent in doing a Deep review on correct and incorrect answers for the Improving and the Dropping groups is shown in Table 5.1. Interesting finding was that such phenomenon was not found during time period of Exam2-Exam3. Students from the Improving group significantly (p=0.02) spent more time reviewing Incorrect questions (M=0.61, SD=0.42) than those who belong to the Dropping group (M=0.42, SD=0.41) during Exam1-Exam2 time period. The results showed that students who improved grades indeed spent more effort on reviewing the problems where they committed mistakes. Unfortunately, such phenomenon was not found in the Exam2-Exam3 time period. Additionally, there were no significant differences found between the two time periods of Improving groups on their Deep Incorrect reviews. This means that students who belong to the Improving group from

| | | Exam1-Exam2 | | Exam2-Exam3 | |
|---|---|---|---|---|---|
| | | Mean | SD | Mean | SD |
| Deep Incorrect | Improving | 0.61** | 0.40 | 0.57 | 0.40 |
| | Dropping | 0.42 | 0.41 | 0.49 | 0.36 |

**Table 5.1:** Deep Incorrect Reviewing of Improving and Dropping Groups in a Formal Assessment ( **p-value<0.01 ; *p<0.05)

the two different time periods actually demonstrated similar reviewing strategies in reviewing incorrect questions. Although there were no significant differences between Improving and Dropping groups during the Exam2-Exam3 period, it could be seen that the Improving Group persistently reviewed incorrect questions to get the wrong right. Meanwhile, the amount of Deep Incorrect reviews increased over time for the Dropping group from time period Exam1-Exam2 to Exam2-Exam3. This explains why there were no significant differences between the Improving and Dropping groups at time period Exam2-Exam3. This also indicates that as grades dropped, students learned to put in more effort in reviewing incorrect questions over time. However, despite the long time they spent on reviewing, their grades did not improve, which suggested that their review actions might be ineffective. The next section discusses the review effectiveness of the Dropping group.

### *5.2.2 Dropping Group Ineffectively Reviewed*

The Dropping group (M=0.17, SD=0.29) was found to have a significantly higher (p=0.002) Deep Correct question review than the Improving group (M=0.03, SD=0.12), refer to Table 5.2. This shows that Dropping group devoted more effort on reviewing correct questions than the Improving group. In Table 5.2, Dropping students spent significantly (p<0.05) longer time in reviewing incorrect questions than correct

| | | Exam1-Exam2 | | Exam2-Exam3 | |
|---|---|---|---|---|---|
| | | Mean | SD | Mean | SD |
| Deep Correct | Improving | 0.03 | 0.12 | 0.07 | 0.18 |
| | Dropping | 0.17** | 0.29 | 0.25** | 0.29 |

**Table 5.2:** Deep Correct Reviewing of Improving and Dropping Groups in a Formal Assessment (**p-value<0.01; *p<0.05)

questions. Dropping students had more incorrect questions than Improving group and they also were considered to be spent more time on incorrect questions than improving group on incorrect question. Surprisingly, the finding was different from my assumption in the previous section. The possible explanation for this phenomenon is that they reviewed correct question to confirm their understanding. The incorrect question may be too difficult for them to understand. Such phenomenon was found in both time period Exam1-Exam2 and Exam2-Exam3. Apparently, this explains the persistent ineffectiveness during the review process. In fact, such inefficient results somehow correspond to the findings in previous HMM analysis (10) students who had difficulties in learning fail to reflect their learning.

## 5.3 Evaluation of Learning Curve

The average time before students review a particular assessment was modelled as a function of review efficiency. Table 4.2 demonstrated the first review attempt review time on exams. This led to an assumption that early review may have a positive influence to changes in the grade, and therefore is considered as an efficient review. Learning curve is introduced to visualize the results. Kullback-Leilber divergence is used to see the efficiency of reviewing strategies of different behavioral groups.

Afterwards, correlation methods were used to explain the impact of reviewing strategy to learning.

### *5.3.1 High Performing Students Were More Vigilant in Review*

Figure 5.1 shows the reviewing learning curve of the High-Level and Low-Level student groups. The x-axis represents the different assessments (exam in black, quiz for attendance in dark blue, and quiz for credit in light blue). The y-axis represents the average time span (in seconds) before the students did the first review. High-Level group is plotted in green, Low-Level group in red, and the average of the two groups in blue. There were several observations based on Figure 5.1: (1) the High-Level students are generally reviewing sooner than Low-Level ones; (2) students tend to review sooner toward the end of the semester; (3) students generally review much sooner in exams than quizzes. However, all these observations were not overly surprising. They suggested that the High-Level students put more effort in reviewing compared to Low-Level students. Additionally, students tend to review exams sooner than quizzes. This could be due to the fact that exams have higher bearing in the course grade.

Based on the above observations, the learning curves of the students, grouped according to their behavior changes, for the two time periods were visualized. The Improving group students attempted to review earlier than the Dropping group. This was especially pronounced in time period the first exam to the second exam. This demonstrated that the Improving groups of students appeared to be more attentive on reviewing their exams/quizzes especially during the Exam1-Exam2 time period. This showed that being more vigilant in reviewing could potentially be associated to the improvement in grades. Thus, in the next section, quantified data on students review efficiency was discussed.


**Figure 5.1:** The Reviewing Learning Curve

### *5.3.2 Improving Group Students Persistently Review; Dropping Group Ineffectively Review*

According to the divergence between groups and divergence between different time periods, the results showed that the Improving group demonstrated a cohesive review strategy between different time periods (no significant differences between time periods) (Table 5.3, column 2). This showed that the students who improved their grades overtime, behaved similarly to persistently review their assessments. In addition, the Improving group also showed significantly higher divergence degree than the Dropping one at Exam1-Exam2 time period ($p<=0.01$) (Table 5.3, row 2). It indicated that the Improving group indeed had a different strategy than the Dropping group. The assumption was that the learning strategy in Improving group was


**Figure 5.2:** Review Learning Curve based on Improving and Dropping Behavior Groups

supposed to have some similarity and have low divergence within group. However, the finding appeared to be contradictory to the assumptionthe Improving group yielded high divergence degree. The coverage analysis was used to explain this pattern. The Improving group (M=9.8, SD=4.4) reviewed more exams and quizzes than the Dropping group (M=6.9, SD=4.3) did. Due to the larger review coverage in the Improving group, not only it demonstrated the students were much more attentive in review, but also had more review diversity and resulted in larger divergence than the Dropping group. Nevertheless, the result of coverage analysis is not enough to the fact caused high divergence in Improving group. Next, considering the active reviewer coverage, whoever reviewed at least 14 out of 16 total exams and quizzes, the results showed that 25% of Improving group students (14/55) regularly and diligently reviewed opposed to the 12% of them in the Dropping group (6/49). Regardless, the amount of

| | Improving group | Dropping group |
|---|---|---|
| Exam1 to Exam2 | **0.64 | 0.38 |
| Exam2 to Exam3 | 0.54 | 0.51 |

**Table 5.3:** Kullback-Leibler Divergence Result for Improving Group and Dropping Group(**p-value<0.01 ; *p<0.05)

students who did regular review was relatively low in either Improving or Dropping group, which explained why the divergence was high. Interestingly. During time period Exam2-Exam3, both groups review patterns became more homogeneous. There were no significant differences between groups (p=0.38) (Table 5.3, row 3).. Besides, the Dropping group significantly increased the frequencies in review, therefore, larger divergence degree (p=0.02) (Table 5.3, column 3). Possible explanations could be that the dropping group students started to worry more about their performances and started to review sooner at the end of the semester. The reviewing behavior was changed to become more active. However, Dropping group reviewed sooner, their grades still decreased. It showed they review in effective way.

### *5.3.3 Review and Learning Impacts*

In section 5.3.2, students review strategies had been found, which the Improving group persistently reviewed and the Dropping group ineffectively reviewed. Next, the relationship between review strategy and grades. In this subsection, correlational analysis was conducted to evaluate the impact of review strategies and students grades. Students time lag for the first review attempt was correlated to their exam scores. The results showed that Improving Group consistently had a negative correlation between the time to attend to review and their exam scores (Table 5.4, row 2). The negative correlation indicated that the longer time a student waited to attend to

| | Exam1 to Exam2 | Exam2 to Exam3 |
|---|---|---|
| Improving group | -0.16 | -0.22 |
| Dropping group | -0.21 | 0.23 |

**Table 5.4:** Correlation on Students First-Attempt-Review and Average Exam Scores

the first review on the exam, the lower the exam grade s/he got. Interestingly, there was a positive correlation found during Exam2-Exam3 time period for the Dropping Group (Table 5.4). The positive correlation showed that the sooner the students began their first review, the lower exam scores they obtained. This suggested that the Dropping group not only attended to review the exams late, but also appeared inefficient review strategy, which may result in no improvements in their grades.

## 5.4 Subjective Evaluation

At the end of the classroom study, a survey was distributed to collect the user experiences on using WPGA. This was announced during the last week of the semester. All students who used the system were invited to answer questionnaire. They also were informed the survey will not affect their grades. Therefore, they can be honest about their answer. A total of 199 respondents answered the survey. In this thesis, only the responses from the students from the Data Structures and Algorithms class were used in this subjective evaluation (74 out of 199 respondents). Figure 5.3 illustrates the responses to some of the survey questions. The whole questionnaire is attached in the appendix section.

### *5.4.1 Learning with WPGA*

Figure 5.3 illustrates that 54.05% (aggregate of 43.24 and 10.81) of the students responded that WPGA were able to help them learn the class material better. On

the contrary, 25.7% were undecided whether the system was able to help them or not. When asked whether they are going to use WPGA to help them in studying for an exam, 57.5% responded positively while 21.9% were uncertain or undecided. Those are promising results since the system was still in development and continued improving features in 2016 Fall semester. However, it also had around 20% of respondents found WPGA cannot help them on learning which impose us there are still have many aspects could improve for students.

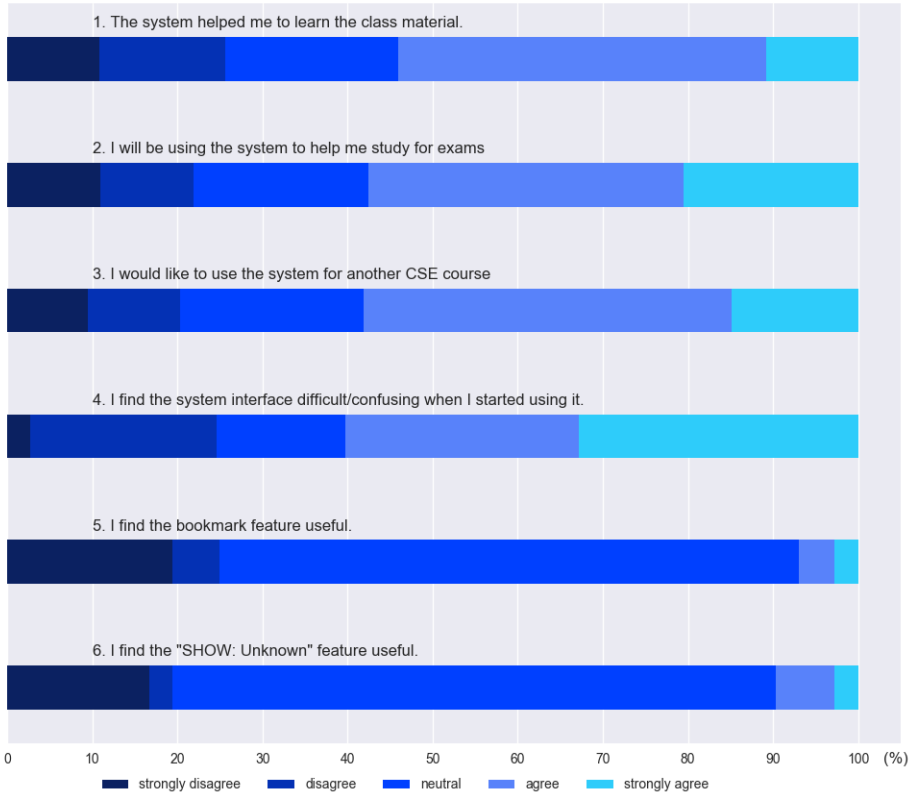


**Figure 5.3:** Part of the Survey Questionnaire Response

### *5.4.2 Ease of Using WPGA*

The survey also told us how difficult for student to use WPGA. In terms of using the WPGA system, 60.3% of the students found it easy to use. Most of the students felt comfortable using the system after taking 1-2 quizzes. It revealed that majority of the users already knew the basic feature like what the color-coding means for a given problem. With WPGA, it is easier for them to access their exam scores virtually anywhere.

**Specific Feature Use**

The students were asked to provide some feedback on some specific features of WPGA. When asked about the usefulness of the bookmark feature, majority (68%) had no opinion about it. This could be attributed to the fact that only few users know it or are aware on how to use it. The same goes for the SHOW: Unknown feature. Majority (70%) had no opinion about it because many of them do not know that the feature exists. The feedback from students encouraged me to add instruction page to guide student to use WPGA.

The following were the features suggested to be included to improve the system: (1) make available to the students the analytics showing the overall performance of everyone in the class; and (2) include social and peer learning features which will allow them to communicate not only with instructors but with other students as well.

# Chapter 6

# CONCLUSIONS AND DISCUSSIONS

## 6.1 Summary

The goal of the project was to study students' learning efforts through their use of Web Programming Grading Assistant (WPGA), a homegrown educational Web application that assists grading and feedback delivery of paper-based assessments. A classroom study was conducted where data from a Data Structure and Algorithms course were collected. Students were grouped according to two criteria: (1) their overall academic performance (Low-Level and High-Level); and (2) behavior change between exams (Improving, Retaining, and Dropping). With first grouping, the High-Level students, on average, significantly had fewer review actions than Low-Level students did. However, High-Level students were able to review more effectively than Low-Level ones, based on the learning curve. Then, research was studied on effectiveness of reviewing behavior on groups of Improving, Retaining, and Dropping respectively. In reviewing formal assessments, student from Improving group focused on their mistakes. This is an efficient learning behavior. In addition, students from Improving group demonstrated their willingness on learning in terms of attending to review promptly and some students in Improving group had persistently reviewing behavior. On the contrary, Dropping students had ineffectively reviews by focusing mostly on the questions they got correctly. They tried to improve their learning strategy after the second exam but no improvements in their grades from correlational analysis. From analysis, students from the Dropping group should focus on their mistakes or misunderstanding to improve their grades.

## 6.2 Contribution

This thesis presents a new educational technology, WPGA. It supports digitizing paper-assessment to cyberspace and has efficient online feedback delivery. A series of user studies were designed and conducted to collect students use of the tool. Moreover, to examine the learning effects and system impacts, students reviewing behaviors were modeled and analyzed. In the preliminary study, students sequential reviews and reflects were modeled (10). Based on the findings, this thesis followed up with deeper analyses in investigating review efficiency and effectiveness. There are several educational implications can be concluded from this work: (1) The high performance students have strategic difference on reviewing; (2) The grade improving students indeed invested time in review and persistently review; (3) The grade dropping students had inefficient reviewing behavior even they spent as same amount time as improving students; (4) Students appreciated WPGA to help them in study, in contingent on system features upgrades. In summary, the research results suggest that students should spend adequate time on reviewing. Additionally, efficient review strategy can involve with persistent reviewing and be more mindful in attending to the exams after they were published.

## 6.3 Limitation and Future Work

There are some limitations in this study. In the analysis, only the students from the Improving group and Dropping group were considered. Those from the Retaining group were not included. Moreover, the Retaining group comprised roughly half of students enrolled in the course. It would be interesting to investigate further the behavior of the students from this group. Furthermore, The study only focused on students voluntarily reviewing as one of the self-regulated learning processes– a

learning behavior of reviewing and evaluating on their own. In the future, a more comprehensive scenario to encourage student learning such as planning, process or comprehension monitoring, and self-explaining should be considered. For system designed, WPGA was introduced to students without a tutorial or a manual. Apart from the class introduction, the only way for the students to be familiarized with the system is to explore by themselves.

The WPGA will be updated based on the feedback from the students. Features such as visualization of learning progress and social features for peer learning. More classrooms studies will be conducted to collect students behavioral data and planning for more exhaustive analyses. The goal overall is to assist students to acquire maximum learning feedback from analytics and the system.

# REFERENCES


[1] Ambrose, S. A., Bridges, M. W., DiPietro, M., Lovett, M. C., & Norman, M. K., How learning works: Seven research-based principles for smart teaching, 2010, John Wiley & Sons.

[2] Bloomfield, A. and Groves, J. F., A tablet-based paper exam grading system, ACM SIGCSE Bulletin, 40(3), 2008, p. 83-87.

[3] Chi, M. T., Self-explaining expository texts: The dual processes of generating inferences and repairing mental models, Advances in Instructional Psychology, 5, 2000, p. 161-238.

[4] Cutumisu, M. and Schwartz, D. L., Choosing versus Receiving Feedback: The Impact of Feedback Valence on Learning in an Assessment Game, The 9th International Conference on Educational Data Mining, 2016, Raleigh, NC.

[5] Dihoff, R. E., Brosvic, G. M., Epstein, M. L., & Cook, M. J., Provision of feedback during preparation for academic testing: Learning is enhanced by immediate but not delayed feedback, The Psychological Record, 54(2), 2004, p. 207-231.

[6] Edwards, S.H. and Perez-Quinones, M. A., Web-CAT: Automatically grading programming assignments, ACM SIGCSE Bulletin, 2008.

[7] Hattie, J. and Timperley, H., The power of feedback, Review of Educational Research, 77(1), 2007, p. 81-112.

[8] Hsiao, I. H., Mobile Grading Paper-based Programming Exams: Automatic Semantic Partial Credit Assignment Approach, The 11th European Conference on Technology Enhanced Education, 2016, Lyon, France: Springers.

[9] Hsiao, I. H., Govindarajan, S. K. P., and Lin, Y. L., Semantic Visual Analytics for Today's Programming Classrooms, The 6th International Learning Analytics & Knowledge Conference, 2016, Edinburgh, UK: ACM.

[10] Hsiao, I. H., Huang, P. K., & Murphy, H., Uncovering reviewing and reflecting behaviors from paper-based formal assessment, Proceedings of the Seventh International Learning Analytics & Knowledge Conference, 2017, p. 319-328. ACM.

[11] Jackson, D. and Usher, M., Grading student programs using ASSYST, ACM SIGCSE Bulletin, 1997.

[12] Kulkarni, C. E., Bernstein, M. S., and Klemmer, S. R., PeerStudio: Rapid peer feedback emphasizes revision and improves performance, Proceedings of the Second ACM Conference on Learning Scale, 2015.

[13] Martinez-Maldonado, R., Dimitriadis, Y., Martinez-Monés, A., Kay, J., & Yacef, K., Capturing and analyzing verbal and physical collaborative learning interactions at an enriched interactive tabletop, International Journal of Computer-Supported Collaborative Learning, 8(4), 2013, p. 455-485.

[14] Murphy, H. E., Digitalizing Paper-Based Exams: An Assessment of Programming Grading Assistant, Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education, p. 775-776.

[15] Paredes, Y.V., Huang, P.K., and Hsiao, I. H., The Role of Reviewing Formal Assessments in Programming Learning, The Thirteenth International Computing Education Research Conference, ACM. (Submitted April 8, 2017, Under review)

[16] Roscoe, R. D. and Chi, M. T., Understanding tutor learning: Knowledge-building

and knowledge-telling in peer tutors' explanations and questions, Review of Educational Research, 77(4), 2007, p. 534-574.

[17] Singh, A., Karayev, S., Gutowski, K., & Abbeel, P., Gradescope: A Fast, Flexible, and Fair System for Scalable Assessment of Handwritten Work, Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale, p. 81-88.

[18] Tempelaar, D. T., Rienties, B., and Giesbers, B., In search for the most informative data for feedback generation: Learning Analytics in a data-rich context, Computers in Human Behavior, 47, 2015, p. 157-167.

[19] Trees, A. R. and Jackson, M. H., The learning environment in clicker classrooms: Student processes of learning and involvement in large university level courses using student response systems, Learning, Media and Technology, 32(1), 2007, p. 21-40.

# APPENDIX A

# QUESTIONNAIRE

1. I would you like to use WPGA for another CSE course.

   ◯ 1: never want to use it at all
   ◯ 2: dont want to use it
   ◯ 3: no opinion
   ◯ 4: would like to use it
   ◯ 5: absolutely would like to use it

2. Rank how helpful WPGA is in helping you to learn the class material.

   ◯ 1: not helpful at all
   ◯ 2: not helpful
   ◯ 3: no opinion
   ◯ 4: helpful
   ◯ 5: absolutely helpful

3. You will be using WPGA to help you study for exams

   ◯ 1: never
   ◯ 2: probably not
   ◯ 3: maybe
   ◯ 4: probably yes
   ◯ 5: definitely yes

4. What do you normally do to study for programming exams that has proved to be the most effective (unrelated to WPGA)?

   ☐ Create a study guide
   ☐ Read/review the textbook
   ☐ Review assignments.
   ☐ Review the slideshows from lecture.
   ☐ Watch free online tutorials/videos.
   ☐ Other

5. Rank how difficult/confusing the WPGA interface was when you first started using it.

   ◯ 1: very difficult
   ◯ 2: somehow difficult
   ◯ 3: no opinion
   ◯ 4: somehow easy
   ◯ 5: very easy

6. Do you know how to use the following features? (check, if you know how to)

   ☐ SHOW: Important
   ☐ SHOW: Bookmark
   ☐ SHOW: Unknown
   ☐ I know what a yellow question means.
   ☐ I know what a green question means.
   ☐ I know what a red question means.
   ☐ I know how to bookmark a question.
   ☐ I know how to view a question.
   ☐ I know how to see the correct answer for a question.
   ☐ I know how to make a note.
   ☐ Other

7. How long did it take you to feel comfortable using WPGA?

   ◯ Right after I log on the system
   ◯ After the 1 2 quizzes.
   ◯ After the 2 3 quiz.
   ◯ After the first test.
   ◯ I still don't feel comfortable using WPGA.

8. Rank how useful you find the bookmark feature.

   ◯ 1: absolutely NOT helpful
   ◯ 2: somewhat helpful
   ◯ 3: no opinion
   ◯ 4: somewhat helpful
   ◯ 5: absolutely helpful

9. Rank how useful you found the "SHOW: Unknown" feature.

   ◯ 1: absolutely NOT helpful
   ◯ 2: somewhat helpful
   ◯ 3: no opinion
   ◯ 4: somewhat helpful
   ◯ 5: absolutely helpful

10. Is there a feature you wish the interface had that it doesn't?
    Ans:

11. I would like the ability to communicate with my professor or TA through WPGA for these reasons:

    - ☐ Rebuttal about points awarded on a specific question on an exam.
    - ☐ Help understanding a specific question.
    - ☐ General questions/communication (I would find it easier to use WPGA than send an email.)
    - ☐ Discuss online with Professor/TA/Peers
    - ☐ I wish WPGA can suggest what "content" that I should focus on
    - ☐ Others: