Computers & Education 82 (2015) 366–377

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

The influences of a two-tier test strategy on student learning: A lag sequential analysis approach

Tzu-Chi Yang [a], Sherry Y. Chen [b,*], Gwo-Jen Hwang [c]

[a] Institute of Information Science, Academia Sinica, Taiwan
[b] Graduate Institute of Network Learning Technology, National Central University, Taiwan
[c] Graduate Institute of Digital Learning and Education, National Taiwan University of Science and Technology, Taiwan

ARTICLE INFO

Article history:
Received 1 July 2014
Received in revised form
18 November 2014
Accepted 26 November 2014
Available online 12 December 2014

Keywords:
Evaluation of CAL systems
Interactive learning environments

ABSTRACT

Recently, programming skills have become a core competence. Many teaching strategies were developed to improve programming skills. Among them, online tests were widely applied to enhance students learning. Nonetheless, they may not be able to engage students in deep thinking and reflections. Thus, a two-tier test strategy was proposed to address this issue. However, previous research mainly focused on investigating the effectiveness of the two-tier test strategy but there is a lack of studies that investigate why the two-tier test approach is effective. To this end, we developed an online test, where the two-tier test strategy was implemented. Additionally, an empirical study was conducted to explore the influences of the two-tier test approach on students' learning performance and behavior patterns. Pre-test and post-test scores were applied to assess students' learning performance while a lag sequential analysis was used to analyze behavior patterns. Regarding learning performance, the proposed two-tier test can improve students' programming skills. Regarding behavior patterns, the two-tier test approach facilitates students to develop a *learning by reviewing* strategy, which is useful to improve their programming skills.

1. Introduction

The rapid development of information technology recently created high demands for software and mobile applications, which were developed with complex programming skills. Therefore, programming skills have become a core competence (Fessakis, Gouli, & Mavroudi, 2013; Hwang, Shadiev, Wang, & Huang, 2012; Verdú et al., 2011), especially for engineering and computer science students (Ala-Mutka, 2005; Brito & Sá-Soares, 2014; Kordaki, 2010). However, learning programming languages is usually challenging and difficult for most students (Brito & Sá-Soares, 2014; Fitzgerald et al., 2008; Rubio-Sánchez, Kinnunen, Pareja-Flores, & Velázquez-Iturbide, 2014). This is due to the fact that the programming languages involve comprehension and practice of a range of theoretical background, semantic and syntactic knowledge, coding skills, and algorithmic skills (Katai & Toth, 2010; Wang, Li, Feng, Jiang, & Liu, 2012). Due to such complexity, many instructors have encountered difficulties in teaching programming languages.

To cope with such difficulties, various teaching strategies and learning activities have been applied to support programming courses. An early study by Machanick (2007) proposed the idea of abstraction-first teaching by hiding details until students are ready for them. Later on, Emurian, Holden, and Abarbanel (2008) employed a peer tutoring approach for improving students' learning performance in a programming course. Subsequently, Hwang et al. (2012) proposed a web-based programming-assisted system, which improved students' program skills by reviewing peers' codes and delivering feedback to peers. More recently, Rubio-Sánchez et al. (2014) implemented an online judge tool, i.e., Mooshak, which provided feedback to help students check whether their computer programs were correct by themselves.

The aforementioned studies suggest that a variety of approaches can be applied to support programming learning. Among them, online tests are a popular activity, which is not only adopted to help students review what they have learnt, but also is employed to examine their learning status (Trotman & Handley, 2008). These are the advantages of online tests. Nonetheless, online tests also have some disadvantages. In particular, conventional online tests request students to provide answers only. On the one hand, students may give the answers with

* Corresponding author.
E-mail addresses: tcyang.academic@gmail.com (T.-C. Yang), sherry@cl.ncu.edu.tw (S.Y. Chen), gjhwang.academic@gmail.com (G.-J. Hwang).

http://dx.doi.org/10.1016/j.compedu.2014.11.021
0360-1315/© 2014 Elsevier Ltd. All rights reserved.

incorrect knowledge (Schuwirth & Van Der Vleuten, 2004). On the other hand, students can guess the answers (Tamir, 1991). For instance, students can provide correct answers by filtering incorrect answers (Kaplan & Saccuzzo, 2001). That is, students may give correct answers not because what they understand is right but because what they guess is right. As highlighted by Schoenfeld (1986, 1988), students usually employ "guess and test loops" as a problem solving strategy in higher education. The problem of such a strategy is that students may not really know why the answers are correct though their answers are correct. In other words, conventional online tests may not be able to engage students in deep thinking and making reflections. Thus, it is worth paying attention to investigating students' thoughts when they interact with the online test.

To address this issue, Treagust (1988) proposed a two-tier test, which can be applied for either evaluating students' knowledge or diagnosing students' misconceptions or alternative conceptions in science Such an approach allows teachers or researchers to not only understand students' possible incorrect ideas, but also to assess their reasoning or understandings behind these ideas. By doing so, suitable support can be provided (e.g., feedback, learning materials). Thus, subsequent researchers (Odom & Barrow, 1995; Tsai, 2001) argued that the two-tier test is an effective way to assess students' knowledge, misconceptions or alternative conceptions. In particular, it is believed that the two-tier test approach can make a great contribution to science education (Chou, Chan, & Wu, 2007; Chu, Hwang, Tsai, & Tseng, 2010; Tsai, 2001).

However, these researchers mainly emphasized on the effectiveness of the two-tier test approach and ignored why the two-tier test approach is effective. To this end, the study presented in this paper attempted to fill this gap by investigating how students use a two-tier test. In other words, the current study aims to illuminate how students react to a two-tier test and a conventional test. To acquire this aim, an empirical study has been conducted to answer the research question – how students with a two-tier test react differently from those with a conventional test.

To answer this research question, a two-tier test strategy is incorporated into an online test in this study, where students' learning performance and behavior patterns were investigated. The former was assessed based on students' pre-test and post-test scores while the latter were analyzed by a lag sequential analysis. Accordingly, the influences of the two-tier test on student learning can be comprehensively investigated. By doing so, we can obtain a deep understanding of why the two-tier test approach is effective.

2. Literature review

2.1. Online tests

In the past decade, various teaching strategies and learning activities have been applied to support computer programming courses (Govender & Grayson, 2008; Kalles, 2008; Machanick, 2007). In particular, many web-based learning environments have been widely applied to train students' programming skills. For example, Emurian et al. (2008) employed a web-based peer tutoring approach for enhancing learning performance of programming languages. Later on, Gálvez, Guzmán, and Conejo (2009) used a web-based problem-solving environment to diagnose students' knowledge levels and to generate feedback and hints to help students correct their misconceptions of programming languages.

A major issue for such popularity lies within the fact that Web-based learning provides many benefits, among which diagnosing students' status and giving feedback have great positive effects. For example, Hwang, Wang, Hwang, Huang, and Huang (2008) developed a web-based learning environment, which could provide guidance based on students' knowledge levels and learning problems that they met. They found that such an approach could improve students' learning performance. Another study by Hwang et al. (2012) proposed a web-based programming-assisted system, named as EduPCR, in which an instant feedback from peers was presented after students submitted their codes. On the other hand, students also played a reviewer to assess peers' codes. The result showed that such feedback could greatly improve students' programming skills because they could know what should be improved by comparing their own source codes with others. Later on, Hauswirth and Adamoli (2013) proposed an online test system which allowed students to learn from their mistakes via answering a series of questions. The students argued that such a learning system was not only useful, but also helped them better understand programming problems. In brief, the aforementioned Web-based learning systems could not only assess students' learning status, but also provided feedback based on their learning status.

Accordingly, assessing students' learning status is essential. To this end, a number of ways were proposed to assess students' learning status. Among them, an online test is a popular activity adopted by teachers (Keppens & Hay, 2008; Trotman & Handley, 2008; Wang et al., 2012). This may be due to the fact that online tests are not only a useful learning activity but also are helpful for students to review what they have learnt (Roediger & Karpicke, 2006). Moreover, they have positive effects on students' learning retention because they can facilitate students to think and practice what they have learnt (Butler, Karpicke, & Roediger, 2007). However, conventional online tests (e.g., multiple-choice questions, simple-answer questions) have some problems. For example, students give correct answers not because what they understand is right but because what they guess is right. Consequently, some researchers acknowledged that a two-tier test may be a promising approach to cope with this issue (Tsai, 2001).

2.2. Two-tier tests

The two-tier test mainly aims at diagnosing students' misconceptions or alternative conceptions. In the first tier, students' descriptive or factual knowledge of the phenomenon is assessed. In the second-tier, students' reasons for their choices made in the first tier are justified. In other words, the two-tier test is to conduct an in-depth investigation of the knowledge that students have learnt. Furthermore, the two-tier test is useful for instructors to have a deep understanding of students' misconceptions because there is a chance to identify why they have such misconceptions (Chu et al., 2010).

Due to such usefulness, research into the two-tier test has mushroomed. Such research indicated that the two-tier test is an efficient and effective way of investigating students' knowledge, misconceptions or alternative conceptions (e.g., Tsai, 2001; Tsai & Chou, 2002). However, such research mainly focused on investigating the effectiveness of the two-tier test strategy. For example, Chou et al. (2007) used a two-tier test to assess students' understandings and alternative conceptions and then provided learning materials. Their results indicated that the two-tier test was useful to enhance students' understandings. They, therefore, addressed that the two-tier test was an effective approach. Other studies, such as Özmen (2008) and Chu et al. (2010), also demonstrated that the two-tier test had positive effects. However, it is still unknown why the two-tier test had such positive effects. In other words, there is a need to develop a deep understanding of the influences of the two-tier test. To develop such a deep understanding, it is necessary to use a new approach to explore students' learning processes in such an online test environment (Cheng & Tsai, 2013).

In this regard, a lag sequential analysis (Bakeman & Gottman, 1997) offers a method that helps researchers examine sequential relationships between each learning behavior based on a statistical theory. Moreover, it allows us to simply identify the significant sequential behavior patterns and to illustrate the relationship of the behaviors with visual diagrams. In recent years, several studies applied lag sequential analysis to do behavioral analysis in educational contexts (Hou, Chang, & Sung, 2007; Lin, Duh, Wang, & Tsai, 2013; Sung, Chang, Lee, & Yu, 2008; Sung, Hou, Liu, & Chang, 2010). In particular, it is appropriative in discovering behavior patterns in online learning environments (Eryilmaz, Chiu, Thoms, Mary, & Kim, 2014; Hou, 2012; Lan, Tsai, Yang, & Hung, 2012). Consequently, a lag sequential analysis is also considered in this study. In brief, this study not only proposes a two-tier test system to discover its impacts, but also uses a lag sequential analysis to identify students' behavior patterns. By doing so, we can obtain a deep understanding of why the proposed two-tier test system has such impacts.

3. Two-tier test-based learning system

In this study, the two-tier test strategy is incorporated into an online test, which was developed to support a programming language course. The proposed Two-Tier Test-based Learning System consists of three subsystems: the On-line Test Subsystem (OTS), the Diagnosis Subsystem (DS) and the Learning Content Interactive Subsystem (LCIS). The OTS provides two online test approaches, i.e., a conventional online test approach and two-tier test approach. When students finish taking a test with the OTS, the DS infers their misconceptions or misunderstandings by judging the answers and reasons provided by them. Subsequently, the LCIS immediately provides corresponding feedback, including the correct answer, an explanation of the answer, and the supplementary material for the students to overcome the misconception. In brief, the OTS offers the online test interaction while the DS serves as a diagnostic tool for evaluating students' learning status. Subsequently, the LCIS provides students with feedback to support their learning. Fig. 1 shows the framework of the two-tier test. Moreover, the design rationale of each subsystem is detailed in subsections below.

3.1. On-line test subsystem

Previous research indicated that online tests offer an efficient assessment to improve teaching and learning (e.g., Hwang, Panjaburee, Triampo, & Shih, 2013). This is due to the fact that students' misconceptions can be identified via online tests and then corresponding feedback is given to address their misconceptions. Such assessment included three activities, i.e. conventional online test, instant feedback, and learning contents. Fig. 2 is an example, which illustrates the process of these activities. As shown in this figure, not only be the correct answers highlighted, but also students are provided with feedback and relevant learning materials. In other words, immediate corresponding feedback is delivered to students after they finish taking a test. The feedback includes a correct answer, an explanation of the answer, and related materials for the misconceptions. Such a process is iterative until students finish taking all of the tests (Hwang, Tseng, & Hwang, 2008; McDaniel, Agarwal, Huelser, McDermott, & Roediger, 2011). Thus, online tests can be considered as a *test-training-test* teaching strategy.

Nevertheless, the aforementioned 'test-training-test' teaching strategy has some problems. For example, students give correct answers not because what they understand is right but what they guess is right. This is due to the fact that such a strategy focused on the answers and ignored students' reasoning processes. To address this problem, we integrate a 'two-tier test' strategy into our online test. Unlike conventional online tests, our two-tier test takes a diagnostic teaching approach. More specifically, a student is provided three or four options and then choose one as his/her answer in the first tier. Likewise, the student is provided three or four options and then choose one to

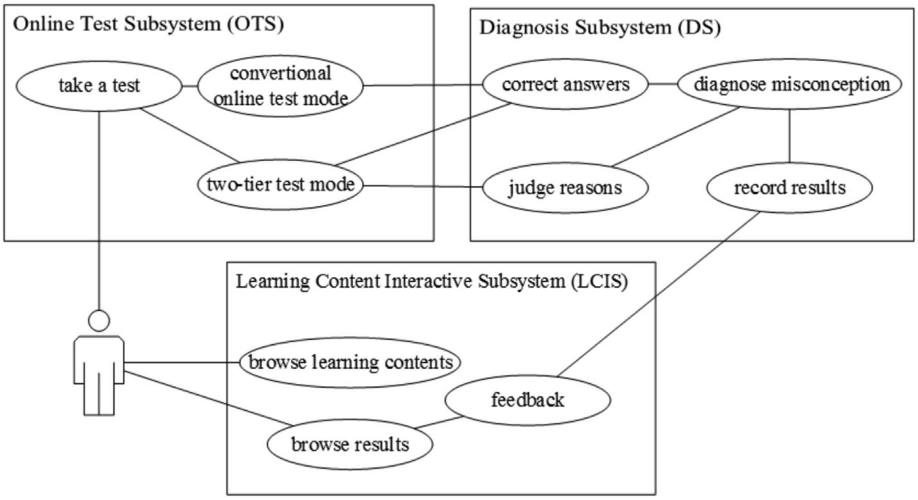
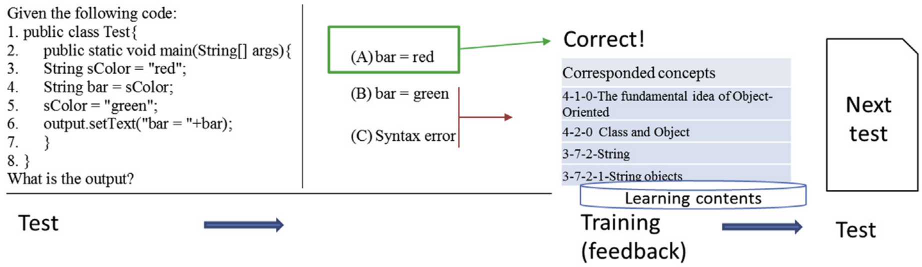


Fig. 1. The framework of the two-tier test based programming language learning system.



```
Given the following code:
public class Test{
    public static void main(String[] args){
    String sColor = "red";
    String bar = sColor;
    sColor = "green";
    output.setText("bar = "+bar);
    }
}
What is the output?
```

Fig. 2. The learning activities of conventional online test.

describe a reason for his/her answer in the second tier. While choosing the reason for his/her answer in the second tier, he/she is allowed to look at the answer given by him/her in the first tier. Fig. 3 shows an example of how the two-tier test diagnoses students' understandings. As shown in this figure, the 'two-tier test' strategy is not only concerned with what answers students give, but also pay attention to examining why they give such answers. By doing so, the two-tier test can diagnose if the students have misunderstandings and provide precise feedback based on reasons that the students describe. In other words, diagnosing students' understandings plays an important role in the two-tier test.

To effectively diagnose students' understandings, there is a need to have a reliable test item bank. Therefore, eight graduate students with more than two-year programming experience were employed to contribute to the development of the test item bank, in which there were 147 two-tier test items, including syntax, variables, type, flow control, and the concepts of Object-Oriented programming. Each test item was reviewed and refined by two teachers who had been teaching a programming course for more than five years.

The test items can be divided into three types: (a) regular quiz: to take place every week to support the weekly course; (b) reviewing test: to include more concepts or units that they have learnt; (c) summative test: to cover all learning materials that teachers have taught. Accordingly, students are allowed to select the type of the test based on their needs. Regardless of the regular quiz, review test, or summative test, students were requested to complete at least one test before a new lecture starts. Additionally, the sequence of the choices and the programming language variables in the test items were changed so as to prevent students from memorizing the answers.

3.2. Diagnosis subsystem

To assess the learning status of the students, the Diagnosis Subsystem (DS) aims to detect the misconceptions, misunderstandings and alternative concepts. As the matter of fact, the first-tier item is used to assess the students' descriptive or factual knowledge of programming languages. The second tier is applied to identify the students' reasons for their choices made in the first tier. By doing so, students' misconceptions or learning problems can be diagnosed. Such a process is analyzed by using decision-tree rules, of which an example is given in Table 1. As shown in this table, students' understandings are divided into four major levels. The level of the understanding that a student belongs to is not only judged based on the answer chosen by him/her, but also is identified by the reason he/she chooses.

3.3. Learning Content Interactive Subsystem

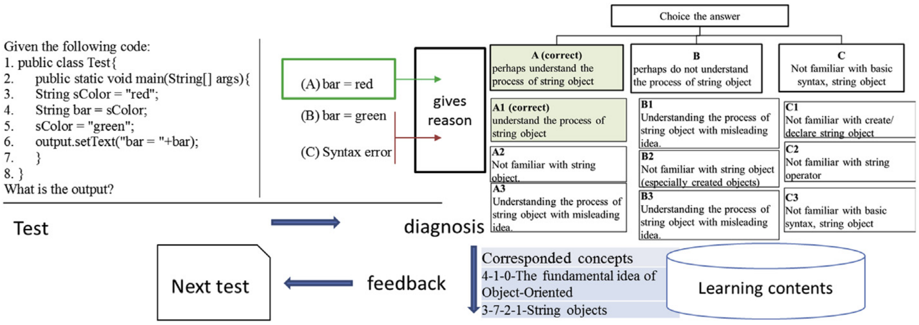
To support the aforementioned teaching strategies, the Learning Content Interactive Subsystem (LCIS) presents current learning contents, including learning materials extracted from the textbook, and demonstrates programming codes provided by the teacher. Students are able to review the contents at any time. According to the students' misconceptions or learning status reported by the DS, the LCIS provides them with relevant learning contents to correct their misconceptions. While students browsing the test result, a list of recommended learning materials was highlighted in the screen and students are suggested to read such materials. Fig. 4 shows an example of providing the learning content based on students' test results. To encourage students to participate in the online learning activity, the leaning contents (e.g., PowerPoint slides, project examples, and supplementary material) are synchronized with those delivered in classrooms.

4. Experiment design

To investigate the effects of the two-tier test strategy on student learning in a programming language course, an experiment was conducted. In addition to comparing the learning performance of students with the conventional test with those with the two-tier test, a lag sequential analysis was employed to identify their different behavior patterns collected from each student's log file. By doing so, we can provide a complete understanding of the effects of the two-tier test, which is one of the contributions of this study. The subsections below describe the details of the experiment, including participants, instruments, and procedures used in the experiment.

4.1. Participants

The experiment was conducted in a "Developing Android applications using JAVA" course delivered for a college in Taiwan. A total of 79 college students participated in this experiment. The average age of the students was 20. All of them had the basic computing and Internet skills necessary to operate the Two-Tier Test-based Learning System. They were randomly divided into an experimental group and a control group. Eventually, 40 students were assigned to the experiment group while 39 students were allocated to the control group.



```
Given the following code:
1. public class Test{
2.      public static void main(String[] args){
3.      String sColor = "red";
4.      String bar = sColor;
5.      sColor = "green";
6.      output.setText("bar = "+bar);
7.      }
8. }
What is the output?
```

(A) bar = red
(B) bar = green
(C) Syntax error
gives reason
Choice the answer
A (correct) perhaps understand the process of string object
B perhaps do not understand the process of string object
C Not familiar with basic syntax, string object
A1 (correct) understand the process of string object
A2 Not familiar with string object.
A3 Understanding the process of string object with misleading idea.
B1 Understanding the process of string object with misleading idea.
B2 Not familiar with string object (especially created objects)
B3 Understanding the process of string object with misleading idea.
C1 Not familiar with create/ declare string object
C2 Not familiar with string operator
C3 Not familiar with basic syntax, string object
Test
diagnosis
Corresponded concepts
4-1-0-The fundamental idea of Object-Oriented
3-7-2-1-String objects
Learning contents
feedback
Next test

Fig. 3. The learning activities of two-tier test approach.

4.2. Instruments

4.2.1. Course material

This course was designed for teaching how to use the Java to develop mobile applications for the Android platform. Because most of the participants do not have experience in writing programs with the JAVA, comprehensive information is covered in this course, including 'introduce the java', 'runtime environment', 'syntax', 'android platform', and 'conduct an android project'. The learning objectives include the understandings of the syntax of JAVA, learning the debugging skills, and programming with the JAVA statements to complete specified projects. To enhance the reliability of the results, all of the students were taught by a same tutor, used the same learning materials and were given same assignments.

4.2.2. Pre-test and post-test

The pre-test and post-test were designed to assess the participants' learning performance both before and after using the proposed Two-Tier Test-based Learning System. Both the pre-test and the post-test were developed by consulting two tutors who had taught the programming language course for more than five years. The pre-test consisted of 33 multiple-choice items with a total score of 100 while the post-test consisted of programming knowledge test and programming skill test with the perfect score 30 and 70, respectively. The programming knowledge (e.g., conceptions, data structure, coding statements) was evaluated with 30 multiple-choice items. The programming skill was assessed by the achievement on solving programming problems (e.g., conducting a small project based on given instructions). The reliabilities of the pre-test and post-test were found to be $r = 0.76$ and $r = 0.77$, respectively.

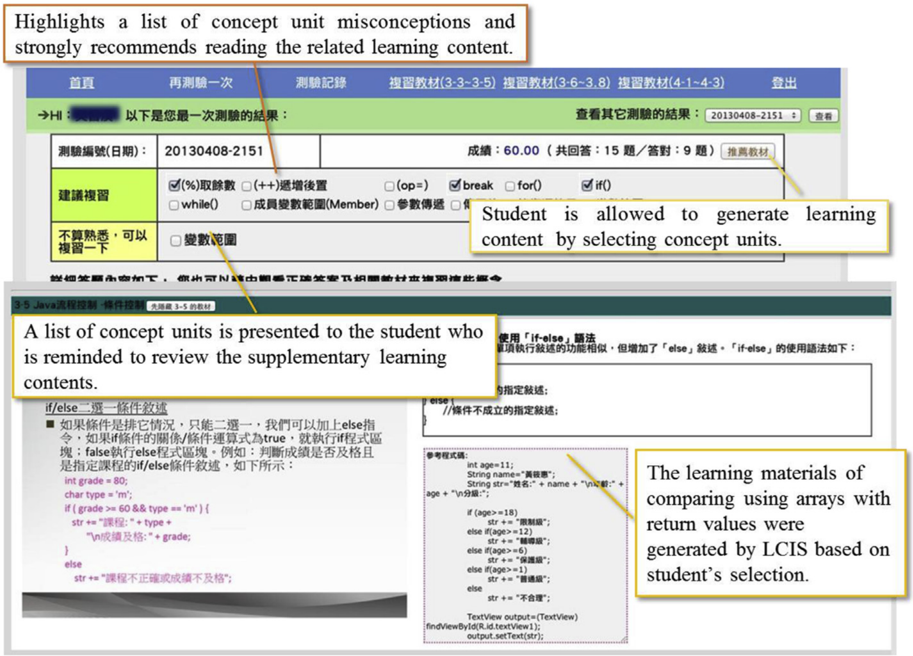
4.2.3. Procedure

In this study, the students in the experimental group learned with the two-tier online test while the students in the control group learned with the conventional online test. Due to such a difference, they received feedback in different ways. More specifically, the students in the experimental group could compare their own reasons with those delivered by the online test. Conversely, such comparison would not be undertaken by the students in the control group. However, both groups were given correct answers and explanations for the rationale of the correct answers. In addition to such information, they were allowed to browse the same learning materials and did the same exercises.

Regardless of the experiment group or the control group, the procedure consists of three stages (Fig. 5). In the first stage, both groups received face-to-face instruction for four weeks. That is, all students received equal instruction on the basic syntax and executive environment of JAVA at the beginning. The students were then asked to complete a segment of code and to perform an exercise based on what they had learned in their classroom. Subsequently, they took a pre-test to analyze their preliminary knowledge of the programming

Table 1
An example of decision-tree rules.

Diagnosis	Rules
(1) Understand the string objects	If the student choose the answer (A) *bar = red*, and gave the reason that (A1) *bar refers to string object "red" via foo*, then the student understands the value assignment of string objects.
(2) Understand, but do not familiar with	If the student choose the answer (A) *bar = red*, and gave the reason that (A2)*"red" assigned to bar immediately*, then the student understands string objects, but perhaps not familiar with it. If the student choose the answer (A) *bar = red*, and gave the reason that (A3) *bar assigned by a copy from foo*, then the student do not understand on the value assignment of string object. If the student choose the answer (B) *bar = green*, and gave the reason that (B2) *bar and foo become the same object since the statement String bar = foo*, then the student do not familiar with the concept of string objects, especially created objects. If the student choose the answer (B) *bar = green*, and gave the reason that (B3) *bar refers to foo*, then the student do not familiar with string objects.
(3) Do not understand the string objects	If the student choose the answer (B) *bar = green*, and gave the reason that (B1) *assigned to the same memory address as foo*, then the student do not understand the value assignment of string objects. If the student choose the answer (C) *Syntax error*, and gave the reason that (C1) *The String should be string (in lower-case)*, then the student do not familiar with declaration of the string objects.
(4) Need to improve the basic knowledge	If the student choose the answer (C) *Syntax error*, and gave the reason that (C1) *Statement "bar=" + bar is not illegal*, then the student do not familiar with string operators. If the student choose the answer (C) *Syntax error*, and gave the reason that (C1) *In line 5, foo = "green" causes an error*, then the student do not familiar with assignment operator, or perhaps have misunderstand on string object.



Highlights a list of concept unit misconceptions and strongly recommends reading the related learning content.

→HI：[redacted] 以下是您最一次測驗的結果：

查看其它測驗的結果：20130408-2151 查看

測驗編號(日期)：	20130408-2151	成績：60.00（共回答：15 題／答對：9 題）推薦教材
建議複習	☑(%)取餘數 □(++)遞增後置 □(op=) ☑break □for() ☑if() □while() □成員變數範圍(Member) □參數傳遞 □	
不算熟悉，可以複習一下	□變數範圍	

Student is allowed to generate learning content by selecting concept units.

3-5 Java流程控制 -條件控制 先隱藏 3-5 的教材

A list of concept units is presented to the student who is reminded to review the supplementary learning contents.

if/else二選一條件敘述

■ 如果條件是排它情況，只能二選一，我們可以加上else指令，如果if條件的關係/條件運算式為true，就執行if程式區塊；false執行else程式區塊。例如：判斷成績是否及格且是指定課程的if/else條件敘述，如下所示：

```
int grade = 80;
char type = 'm';
if ( grade >= 60 && type == 'm' ) {
  str += "課程: " + type +
     "\n成績及格: " + grade;
}
else
  str += "課程不正確或成績不及格";
```

使用「if-else」語法
單項執行敘述的功能相似，但增加了「else」敘述。「if-else」的使用語法如下：

```
的指定敘述;
} else {
    //條件不成立的指定敘述;
}
```

參考程式碼:

```
        int age=11;
        String name="黃筱惠";
        String str="姓名:" + name + "\n年齡:" +
age + "\n分級:";

        if (age>=18)
            str += "限制級";
        else if(age>=12)
            str += "輔導級";
        else if(age>=6)
            str += "保護級";
        else if(age>=1)
            str += "普通級";
        else
            str += "不合理";

        TextView output=(TextView)
findViewById(R.id.textView1);
        output.setText(str);
```

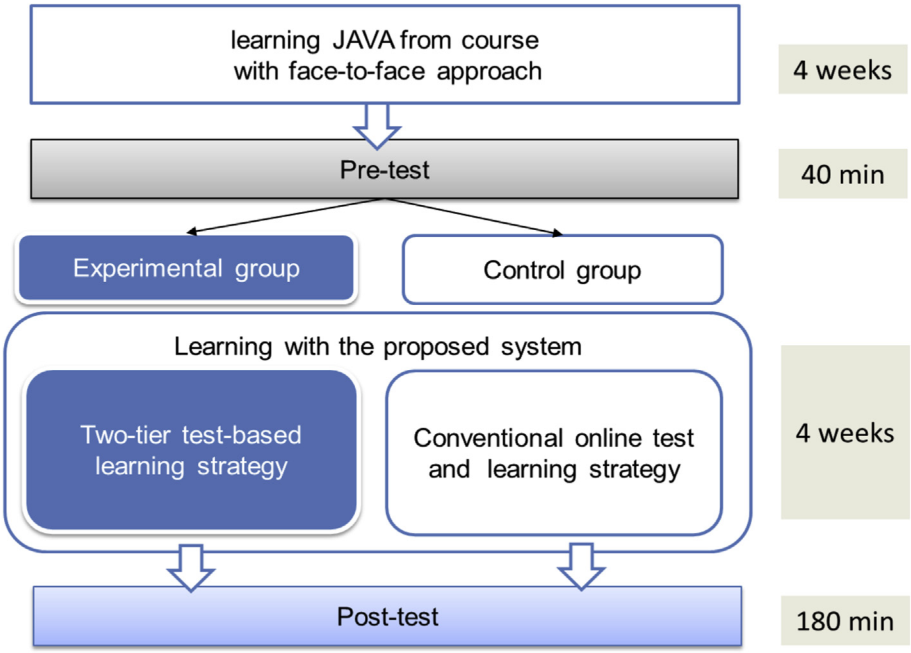
The learning materials of comparing using arrays with return values were generated by LCIS based on student's selection.

Fig. 4. An example of providing feedback and learning materials.

language before interacting with the two-tier online test or conventional online test. In the next stage, both groups were asked to take a test after each lecture was delivered. In general, they would complete at least one test each week. When they interacted with the two-tier online test or conventional online test, their interactions were recorded in a log database. Such interaction lasted for four weeks for both groups. Finally, both groups took a post-test to identify their learning performance.

4.2.4. Data analyses

The independent variable of this study is the online test that students took (i.e., two-tier test or conventional test). The dependent variable is behavior patterns and learning performance. In terms of behavior patterns, a lag sequential analysis (Bakeman & Gottman, 1997) was employed to examine whether the sequential relationships between each behavior pattern reaches statistical significance. Moreover, the lag sequential analysis was also applied to infer the significant behavior patterns and visually represent them. Regarding learning



learning JAVA from course with face-to-face approach
4 weeks
Pre-test
40 min
Experimental group
Control group
Learning with the proposed system
Two-tier test-based learning strategy
Conventional online test and learning strategy
4 weeks
Post-test
180 min

Fig. 5. Experiment procedure.

performance, the pre-test and post-test scores were applied to decide if there are statistically significant differences between two groups. Such analyses were undertaken by the Statistical Package for the Social Sciences (SPSS) and a significance level of $p < 0.05$ was adopted for the study.

5. Results and discussions

5.1. Learning performance

To compare the learning performance of the experiment group and the control group, an ANCOVA was performed on the post-test scores and the pre-test score was considered as the covariant. As shown in Table 2, the experimental group significantly performed better than the control group. As mentioned in Section 4.2.2, the post-test covers the programming knowledge test and programming skill test. Our results indicated that these two groups significantly performed differently in the programming skill test while they demonstrated similar performance in the programming knowledge test (Table 3). These findings suggest that the proposed two-tier test approach has significant effects on programming skills, instead of programming knowledge. These results are supported by previous studies (Chu et al., 2010; Panjaburee, Triampo, Hwang, Chuedoung, & Triampo, 2013). However, it is unclear why such an approach is useful. Accordingly, there is a need to investigate how students behaved differently in the two-tier test and conventional test. Thus, we conduct detailed analyses for their behavior patterns, of which the details are described in the following sections.

5.2. Behavior frequency analysis

The individual behaviors of the 79 participants were coded using the coding scheme, which included seven kinds of codes (Table 4). According to the rationale of the lag sequential analysis, participants' behaviors were coded in the chronological order of their occurrences. For example, after logging into the system (H), a student takes a regular quiz (Q1), then the student clicks the button to browse the result of the quiz (R), and selects a part of learning material to read (L); this series of behaviors was thus coded as H Q1 R L. In total, there are 3626 students' behavior codes. Table 5 presents the frequencies of these codes based on the aforementioned seven kinds of code. As shown in this table, the volume of codes in the experimental group (1662 codes) was fewer than that of the control group (1964 codes). This finding suggests that the students in the experimental group seemed to demonstrate fewer learning activities than those in the control group.

According to a Chi–Square test result ($\chi^2(6) = 164.85$, $p < 0.001$), the behavior distribution of the experimental group significantly differ from that of the control group. In terms of the experimental group, C (comparing the results of past quizzes, 544) has the highest frequency followed by H (logging/main page, 381) (Table 5). Likewise, C also has the highest frequency in the control group but the second highest frequency is Q1 (taking a regular quiz, 402). These findings suggested that both groups were concerned with the results of their past quizzes, including the answer to the quiz, explanation for test items, unfamiliar topics, and suggested learning materials. However, a difference exists between these two groups. More specifically, reviewing their past quiz results is a major activity for the experiment group while the control group not only paid attention to comparing their previous quiz results, but also dedicate themselves to taking a regular quiz. In other words, taking a regular quiz is also an important activity for the control group. Thus, these findings suggested that the students with the two-tier test strategy focused on reviewing what they had learnt by comparing the previous quizzes that they took while those with the conventional test strategy put effort to review their past quizzes results as well as to take the tests.

5.3. Behavior sequential analysis

According to the procedures of sequential analysis, we calculated the frequency of each behavioral category immediately following another behavioral category. The results of the experimental group and the control group are shown in Tables 6 and 7, respectively. The first column presents the starting behaviors and the first row describes the behaviors that occurred after the starting behaviors finished. The numbers represent the total number of times a column behavior occurred immediately after a row behavior ended. For example, the frequencies of starting behavior H followed by Q1 is 90 and the frequencies of starting behavior Q1 followed by H is 81.

Based on the frequency transition tables (Tables 6 and 7), we then conducted sequential analysis and further determined whether the connection between each sequence reached statistical significance. The z-score value of each sequence was calculated to determine whether the continuity of each reached the level of significance and a z-value greater than $+1.96$ indicates that a sequence reaches the level of significance ($p < 0.05$). In that case, the codes obtained from the experimental group and control group yielded the adjusted residuals tables (Tables 8 and 9). Furthermore we deduced the behavior-transfer diagrams of the experimental group and the control group, which are presented in Figs. 6 and 7, respectively. These two figures illustrate all sequences that have reached significance and the numerical values in the figures are the sequences' z-scores and the arrow indicates the direction of transfer for each sequence.

5.4. Learning behavior patterns

After comparing the behavioral diagram of the experiment group with that of the control group, we found that there were some similarities between the experiment group and the control group but some differences also existed between these two groups. The details are discussed subsections below.

Table 2
ANCOVA result of the post-test scores of the two groups.

Group	*N*	Mean	S.D.	Adjusted mean	Std. error.	*F*
Experimental group	40	75.54	22.28	75.22	2.85	4.84*
Control group	39	65.97	17.11	66.30	2.88	

*p < .05.

Table 3
The *t*-test results of the programming knowledge and programming skills tests on the two groups.

	Group	*N*	Mean	S.D.	*t*
Programming knowledge test	Experimental group	40	20.49	2.95	−1.442
	Control group	39	21.48	3.12	
Programming skills test	Experimental group	40	55.05	21.00	2.546*
	Control group	39	44.49	15.36	

*$p < 0.05$.

Table 4
Coding scheme of web-based test learning behaviors.

Behavior	Codes	Description
Logging in/main page	H	The learner logging on into the system
Take a regular quiz	Q1	Complete a quiz that the scope of the quiz are current course progress.
Take a reviewing quiz	Q2	The learner takes a quiz which they finished before.
Take a summative test	Q3	The learner completes a quiz that does not limit by specific scope. Every concept which they have learned could be present in this quiz.
Reading learning materials	L	The learner clicks the items for pop-up learning materials and read them.
Review a quiz result	R	Review a result of a quiz.
Compare prior quiz results	C	Compare the result of previous records.

5.4.1. Similarities

These two groups demonstrated similar behavior sequences in taking quizzes and reviewing results, i.e., H ↔ Q1, H → R, Q1 → Q1, Q2 → Q2, Q3 → Q3, L → L, C → C, R → C, where '→'indicates a unidirectional sequence and '↔' indicates bio-directionality. Such behavior sequences demonstrate some interesting trends.

- *Taking regular quizzes and browsing the results are major purposes of taking online tests.* The sequences, H (logging/main page) ↔ Q1 (take a regular quiz), indicated that students would take a regular quiz immediately after they logged in the system and then went back to the main page after completing the quiz without involving another activity. These results imply that taking regular quizzes is a major purpose when students take the online test. On the other hand, the sequences, H → R and R → C (compare past quiz results), indicate that students focused on viewing and comparing the results of the quizzes. In other words, browsing the results is another major purpose for the students to take the online test.
- *Students are keen to take quizzes*. Three types of quiz, Q1, Q2 and Q3, were available in the online test. More specifically, Q1 means that a student takes a regular quiz, Q2 means that a student re-takes a quiz which he/she has taken before, apart from regular quizzes, and Q3 means the student takes a summative test, of which the scope covered all concepts he/she had learnt. As shown in Figs. 1 and 2, students in both groups tended to repeatedly take the regular quiz (Q1 → Q1), to re-take the same quiz (Q2 → Q2) and the summative test (Q3 → Q3). These identical sequential patterns indicated that students were willing to take quizzes, regardless of the conventional test or two-tier test.
- *Students expected to obtain high scores.* Q1 → Q1 showed that some students tended to take the regular quiz repeatedly. Such repetition may be caused by the fact that the students were asked to take a test after each lecture and they were told that the highest score obtained from the quiz was recorded. Accordingly, the students expected to get a high score so they took the quiz several times. This

Table 5
The frequencies of code behaviors in the experimental group and control group.

		H	Q1	Q2	Q3	L	R	C	Totals
Experimental group	Mean	9.97	5.32	2.42	2.32	5.15	5.54	15.32	
	SD	7.79	6.21	6.09	4.19	6.70	5.76	23.05	
	Range	1–30	1–31	0–35	0–17	0–23	0–22	0–102	
	Total	381	204	88	87	193	204	544	1701
Control group	Mean	8.12	11.07	1.98	6.27	5.37	3.78	12.44	
	SD	5.40	10.24	3.13	14.97	4.79	3.76	20.89	
	Range	1–19	1–36	0–11	0–85	0–14	0–15	0–98	
	Total	238	402	78	266	224	166	530	2004

Table 6
Results of frequency transition in the experimental group.

	H	Q1	Q2	Q3	L	R	C
H	89	90	10	26	15	140	3
Q1	81	64	9	3	9	14	15
Q2	17	4	47	2	3	6	9
Q3	21	3	1	40	6	7	4
L	15	11	3	4	122	15	21
R	34	11	3	3	19	12	119
C	85	21	15	9	19	10	373

Table 7
Results of frequency transition in the control group.

	H	Q1	Q2	Q3	L	R	C
H	26	119	13	42	22	105	6
Q1	113	211	5	10	19	9	30
Q2	16	4	38	3	8	4	2
Q3	31	9	4	175	6	5	22
L	30	17	15	7	115	22	16
R	14	7	2	2	42	8	91
C	68	35	1	27	12	13	363

Table 8
The result of the sequential analysis of behavior demonstrated by students in the experimental group. The measurement of significance used in the lag sequential analysis is based on a single z-value. More specifically, the value that is greater than +1.96 is considered as a significant value. Thus, only one * is applied to mark the significance for each entry.

	H	Q1	Q2	Q3	L	R	C
H	1.73	7.9*	−2.56	1.7	−5.19	16.84*	−14.85
Q1	7.93*	9.08*	−0.52	−2.52	−3.33	−2.4	−8.04
Q2	−0.71	−2.21	20.98*	−1.24	−2.41	−1.53	−4.49
Q3	1.72	−2.52	−1.74	17.76*	−1.34	−1.16	−5.62
L	−4.81	−2.86	−2.41	−2.04	24.13*	−1.92	−6.68
R	−1.56	−3.09	−2.55	−2.52	−0.98	−2.86	8.6*
C	−3.2	−7.06	−3.07	−4.43	−6.99	−8.82	22.23*

finding suggested that either the experimental group or the control group followed the course instruction and expected to get a high score from the online test. In other word, the two-tier test strategy does not influence students' intention to use such an online test.

5.4.2. *Differences*

The students in the control group demonstrated significant behavior sequences, i.e., R → L and L → Q2, which are coherent with the test-training-test strategy frequently used in the conventional online tests. More specifically, students in the control group used to browse learning contents after they reviewed the result of a quiz (i.e., R → L). Moreover, the behavior sequence, L → Q2, indicated that they would re-take a quiz with the same scope after browsing the quiz result and learning materials. Stated in another way, they continually reviewed the results, browsing leaning materials and re-taking a quiz. In other words, their learning strategies were to 'take a quiz to find ambiguous understandings of concepts', 'to learn from learning materials' and 'to confirm their understandings by answering questions'. This may be because students were provided with various feedback (e.g., correct answer, highlighted mistakes, and related concepts), which stimulated them to review relevant learning materials. However, they might not have enough confidence so they needed to re-take the quiz to confirm what they have learned after viewing the learning materials.

On the contrary, the experimental group did not reveal such behavior sequences (i.e., R → L → Q2). After taking a further look at the frequency distribution of code behaviors in the experimental group (Table 5), we found that they demonstrated fewer behavior codes on taking tests. These findings suggest that students in the control groups and those in the experimental group demonstrated different behavior patterns. As mentioned in Section 4.1, the experimental group showed better programming skills than the control group. The former took the two-tier test approach while the latter used the conventional test approach. These findings not only suggest that the two-tier test has effects on student's behavior patterns, but also reveal that the two-tier test approach is useful for students to develop behavior patterns that can improve their programming skills.

5.5. *The influence of the two-tier test approach*

In order to further investigate the influences of the proposed two-test, the transition of students' behavior were analyzed. As suggested by Reed and Oughton (1997), students' learning behavior can be divided into three stages, i.e., early, middle and late. In the light of their suggestion, students' learning behavior is also divided into three stages based on the following rules: (a) the early stage: Day 1 to Day 8, (b) the middle stage: Day 9 to Day 19 and (c) the late stage: Day 20 to Day 28. Fig. 8 illustrates the behavioral transition of the experimental group through these three stages while that of the control group is presented in Fig. 9. Furthermore, such behavioral transition is further discussed in subsections below.

Table 9
The result of the sequential analysis of behavior demonstrated by students in the control group. The measurement of significance used in the lag sequential analysis is based on a single z-value. More specifically, the value that is greater than +1.96 is considered as a significant value. Thus, only one * is applied to mark the significance for each entry.

	H	Q1	Q2	Q3	L	R	C
H	−4.01	7.59*	−0.06	−0.53	−3.01	16.63*	−11.3
Q1	7.48*	18.16*	−3.07	−7.13	−4.59	−4.92	−9.65
Q2	1.56	−3.32	21.03*	−2.47	−0.22	−1	−4.84
Q3	0.02	−7.29	−2.16	27.11*	−4.96	−4.07	−7.22
L	−1.09	−4.95	2.3*	−4.75	20.24*	0.89	−6.95
R	−3.03	−5.32	−1.87	−4.79	6.03*	−1.69	8.65*
C	−1.41	−9.02	−5.14	−6.47	−7.59	−5.68	25.59*

$^*p < .05$.

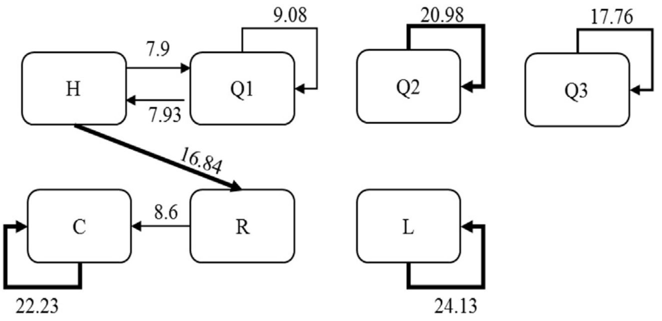


Fig. 6. The behavioral transition diagram of experimental group.

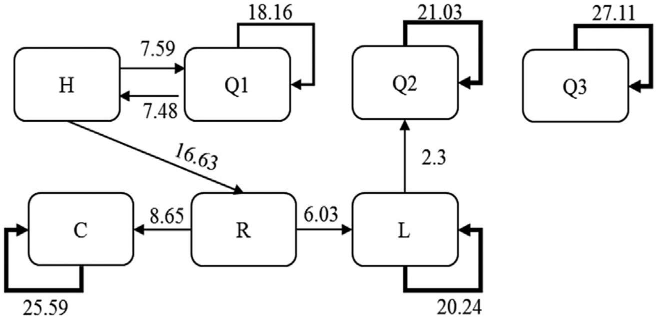


Fig. 7. The behavioral transition diagram of control group.

5.5.1. Early stage

As shown in Figs. 8 and 9, the experiment group and the control group demonstrated similar behavior patterns in the early stage. In particular, both groups tended to take a regular quiz (Q1 → Q1) repeatedly. Such a finding suggests that the proposed two-tier has minor effects on student behavior in the early stage. However, a small difference exists between these two groups. More specifically, the experimental group repeated to take the past quizzes that they had taken before (Q2 → Q2) while the control group did not demonstrate such a behavior pattern. This finding implies that the two-tier test could motivate students to take quizzes in the early stage, especially past quizzes.

5.5.2. Middle stage

In the middle stage, Q2 → Q2 and Q3 → Q3 appear in both groups. It suggested that taking quizzes was a main learning activity, regardless of the test approach. However, a 'test-learning-test' strategy (i.e., R → L → Q2) was found in the control group while a 'learning by reviewing' strategy (i.e., R → C; L → L; Q2 → Q2) was demonstrated in the experimental group. These findings are in line with those presented in Section 5.4.2, which indicated that the proposed two-tier test has effects on students' behavior patterns.

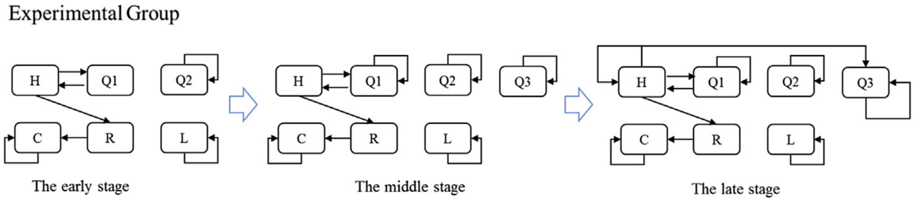


Fig. 8. Sequential-analysis transition diagrams of experimental group in each stage.

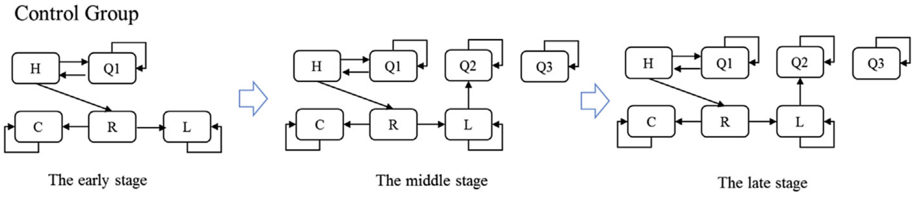


Fig. 9. Sequential-analysis transition diagrams of control group in each stage.

Table 10
Two-tier test approach vs. conventional test approach.

		Two-tier test approach	Conventional test approach
Learning performance		Programming skills improved	None
Learning activity		To review and compare results	To brows learning content
Learning strategies	Early stage	To repeat to take the regular quizzes and past quizzes	To repeat to take regular quizzes only
	Middle stage	To develop a 'learning by reviewing' strategy	To develop a 'test-learning-test' strategy
	Late stage	To move to do a comprehensive review	To keep taking the 'test-learning-test' strategy

5.5.3. Late stage

In the late stage, most students in the control group keep taking the test-learning-test strategy used in the middle stage. Conversely, the experiment group demonstrated a significant behavior sequence (i.e., H → Q3, H → H). More specifically, most students in the experimental group tend to immediately take a summative test (H → Q3). Otherwise, they would logoff the system or go back to the main page (H → H). In other word, the experiment group was keen to take the summative tests and to browse the main page in this stage. The summative tests, which cover a wide range of information, can help students do a comprehensive review. Such findings suggested that student with the two-tier test approach paid attention to reviewing all of the materials, instead of taking the regular quizzes, in the last stage.

In brief, the control group and experiment group demonstrate different behavior patterns during these three stages. These findings echoes those mentioned in Sections 5.1–5.4, students with the two-tier approach and those with the conventional test approach behaved very differently. Table 10 summarizes differences between these two groups, including learning performance, learning objective, learning activity, and learning strategies used in the early, middle and late stages.

6. Conclusions

This study investigates "*how students with a two-tier test react differently from those with a conventional test?*". To answer this research question, both learning performance and behavior patterns are considered. Regarding learning performance, our results show that students' programming skills are improved. Accordingly, these findings imply that the two-tier test strategy is helpful for students to learn a programming language. Regarding behavior patterns, the lag sequential analysis approach was applied to analyze students' behavior patterns. The findings from this study revealed that the students with the two-tier test approach and those with the conventional test showed some similar behavior but there were some differences between them. Regarding the similarities, both groups were keen to take quizzes and expected to obtain high scores. Regarding the differences, students with the conventional test approach developed a *test-learning-test* strategy while those with the two-tier test approach developed a *learning by reviewing* strategy. Such findings reveal that reviewing mechanisms provided by the two-test test play an important role and should be taken into account in the development of future online tests. In general, the findings from this study echoes those from previous research (e.g., Brito & Sá-Soares, 2014), which indicated that the two-tier test approach can help students develop their learning strategies and make proper adjustments.

The contribution of this study includes three aspects: theory, methodology, and applications. In terms of theory, this study deepens the understandings of the impacts of the two-tier test on student learning by providing empirical evidence. The findings of this study indicated that the two-tier test was helpful for students to develop learning strategies that could improve their programming skills. However, it was only one relatively small study. Further work needs to be undertaken with a larger sample to provide additional evidence. With regard to methodology, this study analyzed the experimental data with a lag sequential analysis approach. Such an approach is useful to identify valuable relationships. Given any dataset, there are often no strict rules that impose the use of a specific method over another in its analysis. Therefore, it is necessary to conduct further works to analyze students' behavior patterns with other approaches, e.g., a data mining approach. It would be interesting to see whether similar results can be found by using these approaches.

As far as the application is concerned, this study illustrates how to implement a two-tier online test in web-based learning. Moreover, the results indicated that the proposed two-tier online test can help students developed deep thinking and adjust their learning strategy. In other words, the design approaches used in the proposed two-tier online test seem useful to the practical world but they were applied to support programming courses only in this study. On the other hand, such design approaches can also be implemented to support various courses in the future, such as mathematics, natural science and English.

Fruitful results were obtained from this study but we did not examine how individual differences (e.g., gender, prior knowledge) affect students' reactions to the two-tier online test. This issue should be addressed in our future works. The findings from such future works can be integrated with those from this present work. By doing so, we can develop personalized two-tier tests that can accommodate students' individual differences.

Acknowledgment

The authors would like to thank the National Science Council of the Republic of China, Taiwan, for financial support (NSC 101-2511-S-008 -010 -MY3; NSC 102-2511-S-008 -002-MY2).

References

Ala-Mutka, K. M. (2005). A survey of automated assessment approaches for programming assignments. *Computer Science Education, 15*(2), 83–102.
Bakeman, R., & Gottman, J. M. (1997). *Observing interaction: An introduction to sequential analysis* (2nd ed.). UK: Cambridge University Press.
Brito, M. A., & Sá-Soares, F. (2014). Assessment frequency in introductory computer programming disciplines. *Computers in Human Behavior, 30*, 623–628.
Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied, 13*, 273–281.

Cheng, K. H., & Tsai, C. C. (2013). Affordances of augmented reality in science learning: suggestions for future research. *Journal of Science Education and Technology, 22*, 449–462.

Chou, C., Chan, P. S., & Wu, H. C. (2007). Using a two-tier test to assess students' understanding and alternative conceptions of cyber copyright laws. *British Journal of Educational Technology, 38*(6), 1072–1084.

Chu, H. C., Hwang, G. J., Tsai, C. C., & Tseng, J. C. R. (2010). A two-tier test approach to developing location-aware mobile learning systems for natural science courses. *Computers & Education, 55*, 1618–1627.

Emurian, H. H., Holden, H. K., & Abarbanel, R. A. (2008). Managing programmed instruction and collaborative peer tutoring in the classroom: applications in teaching Java™. *Computers in Human Behavior, 24*, 576–614.

Eryilmaz, E., Chiu, M. M., Thoms, B., Mary, J., & Kim, R. (2014). Design and evaluation of instructor-based and peer-oriented attention guidance functionalities in an open source anchored discussion system. *Computers & Education, 71*, 303–321.

Fessakis, G., Gouli, E., & Mavroudi, E. (2013). Problem solving by 5–6 years old kindergarten children in a computer programming environment: a case study. *Computers & Education, 63*, 87–97.

Fitzgerald, S., Lewandowski, G., McCauleyc, R., Murphyd, L., Simone, B., Thomasf, L., et al. (2008). Debugging: finding, fixing and flailing, a multi-institutional study of novice debuggers. *Computer Science Education, 18*(2), 93–116.

Gálvez, J., Guzmán, E., & Conejo, R. (2009). A blended E-learning experience in a course of object oriented programming fundamentals. *Knowledge-Based Systems, 22*, 279–328.

Govender, I., & Grayson, D. J. (2008). Pre-service and in-service teachers' experiences of learning to program in an object-oriented language. *Computers & Education, 51*, 874–885.

Hauswirth, M., & Adamoli, A. (2013). Teaching Java programming with the Informa clicker system. *Science of Computer Programming, 78*, 499–520.

Hou, H.-T. (2012). Analyzing the learning process of an online role-playing discussion activity. *Educational Technology & Society, 15*(1), 211–222.

Hou, H. T., Chang, K. E., & Sung, Y. T. (2007). An analysis of peer assessment online discussions within a course that uses project-based learning. *Interactive Learning Environments, 15*(3), 237–251.

Hwang, G. J., Panjaburee, P., Triampo, W., & Shih, B. Y. (2013). A group decision approach to developing concept-effect models for diagnosing student learning problems in mathematics. *British Journal of Educational Technology, 44*(3), 453–468.

Hwang, W. Y., Shadiev, S., Wang, C. Y., & Huang, Z. H. (2012). A pilot study of cooperative programming learning behavior and its relationship with students' learning performance. *Computers & Education, 58*, 1267–1281.

Hwang, G. J., Tseng, J. C., & Hwang, G. H. (2008). Diagnosing student learning problems based on historical assessment records. *Innovations in Education and Teaching International, 45*(1), 77–89.

Hwang, W. Y., Wang, C. Y., Hwang, G. J., Huang, Y. M., & Huang, S. (2008). A web-based programming learning environment to support cognitive development. *Interacting with Computers, 20*, 524–534.

Kalles, D. (2008). Students working for students on programming courses. *Computers & Education, 50*, 91–97.

Kaplan, R. M., & Saccuzzo, D. P. (2001). *Psychological testing: Principles, applications, and issues*.

Katai, T., & Toth, L. (2010). Technologically and artistically enhanced multi-sensory computer-programming education. *Teaching and Teacher Education, 26*, 244–251.

Keppens, J., & Hay, D. (2008). Concept map assessment for teaching computer programming. *Computer Science Education, 18*(1), 31–42.

Kordaki, M. (2010). A drawing and multi-representational computer environment for beginner learning of programming using C: design and pilot formative evaluation. *Computers & Education, 54*, 69–87.

Lan, Y. F., Tsai, P. W., Yang, S. H., & Hung, C. L. (2012). Comparing the social knowledge construction behavioral patterns of problem-based online asynchronous discussion in e/m-learning environments. *Computers & Education, 59*, 1122–1135.

Lin, T. J., Duh, H. B. L., Wang, H. Y., & Tsai, C. C. (2013). An investigation of learners' collaborative knowledge construction performances and behavior patterns in an augmented reality simulation system. *Computers & Education, 68*, 314–321.

Machanick, M. (2007). Teaching Java backwards. *Computers & Education, 48*, 396–408.

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L., III (2011). Test-enhanced learning in a middle school science classroom: the effects of quiz frequency and placement. *Journal of Educational Psychology, 103*(2), 399–414.

Odom, A. L., & Barrow, L. H. (1995). The development and application of a two-tiered diagnostic test measuring college biology students' understanding of diffusion and osmosis following a course of instruction. *Journal of Research in Science Teaching, 32*(1), 45–61.

Özmen, H. (2008). The influence of computer-assisted instruction on students' conceptual understanding of chemical bonding and attitude toward chemistry: a case for Turkey. *Computers & Education, 51*, 423–438.

Panjaburee, P., Triampo, W., Hwang, G. J., Chuedoung, M., & Triampo, D. (2013). Development of a diagnostic and remedial learning system based on an enhanced concept effect model. *Innovations in Education and Teaching International, 50*(1), 72–84.

Reed, W. M., & Oughton, J. M. (1997). Computer experience and interval-based hypermedia navigation. *Journal of Research on Computing in Education, 30*(1), 38–52.

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249–255.

Rubio-Sánchez, M., Kinnunen, P., Pareja-Flores, C., & Velázquez-Iturbide, Á. (2014). Student perception and usage of an automated programming assessment tool. *Science of Computer Programming, 31*, 453–460.

Schoenfeld, A. (1986). On having and using geometric knowledge. In J. Hiebert (Ed.), *Conceptual and procedural knowledge: The case of mathematics* (pp. 225–264). Hillsdale, NJ: Lawrence Erlbaum.

Schoenfeld, A. (1988). When good teaching leads to bad results: the disasters of “well-taught” mathematics courses. *Educational Psychologist, 23*(2), 145–166.

Schuwirth, L. W., & Van Der Vleuten, C. P. (2004). Different written assessment methods: what can be said about their strengths and weaknesses? *Medical Education, 38*(9), 974–979.

Sung, Y. T., Chang, K. E., Lee, Y. H., & Yu, W. C. (2008). Effects of a mobile electronic guidebook on visitors attention and visiting behaviors. *Educational Technology and Society, 11*(2), 67–80.

Sung, Y. T., Hou, H. T., Liu, C. K., & Chang, K. E. (2010). Mobile guide system using problem-solving strategy for museum learning: a sequential learning behavioural pattern analysis. *Journal of Computer Assisted Learning, 26*, 106–115.

Tamir, P. (1991). Multiple choice items: how to gain the most out of them. *Biochemical Education, 19*(4), 188–192.

Treagust, D. F. (1988). Development and use of diagnostic tests to evaluate students misconceptions in science. *International Journal of Science Education, 10*(2), 159–169.

Trotman, A., & Handley, C. (2008). Programming contest strategy. *Computers & Education, 50*(3), 821–837.

Tsai, C. C. (2001). The interpretation construction design model for teaching science and its applications to internet-based instruction in Taiwan. *International Journal of Educational Development, 21*, 401–415.

Tsai, C. C., & Chou, C. (2002). Diagnosing students' alternative conceptions in science. *Journal of Computer Assisted Learning, 18*, 157–165.

Verdú, E., Regueras, L. M., Verdú, M. J., Leal, P. J., Castro, J. P., & Queirós, R. (2011). A distributed system for learning programming on-line. *Computers & Education, 58*, 1–10.

Wang, Y., Li, H., Feng, Y., Jiang, Y., & Liu, Y. (2012). Assessment of programming language learning based on peer code review model: Implementation and experience report. *Computers & Education, 59*, 412–422.