To Implement Computerized Adaptive Testing by Automatically Adjusting Item Difficulty Index on Adaptive English Learning Platform

Shu-Chen Cheng[1], Yu-Ping Cheng[2], Yueh-Min Huang[2]

[1] Department of Computer Science and Information Engineering, Southern Taiwan University of Science and Technology, Taiwan

[2] Department of Engineering Science, National Cheng Kung University, Taiwan

kittyc@stust.edu.tw, n98061513@gs.ncku.edu.tw, huang@mail.ncku.edu.tw

Abstract

In previous language teaching studies, there were still some issues that have rarely been discussed or been of concern. One of the issues is the limitation to provide the appropriate test items according to the degree of the students' capability in computerized adaptive testing. However, the difficulty level of a test item relies on expert judgment, which it is too time-consuming and expert-dependent. In order to effectively achieve automatic adjustment of the difficulty levels of test items to implement computerized adaptive testing, an attempt is made to address this issue by developing an adaptive English learning platform. This platform was used to analyze the difficulty levels of the items along with 162 individual ability values from the massive data of the 53,191 item answer records. The research results showed that analyzing the records using big data analysis and the item response theory enables the platform to automatically determine and adjust the difficulty levels of the items and to provide the fitting test items that match learner capabilities in computerized adaptive testing. Therefore, teachers do not have to manually adjust the difficulty levels of the test items. In addition, students can use computerized adaptive tests to improve their learning performance.

Keywords: Item difficulty Index, Item response theory, Computerized adaptive testing

1 Introduction

Currently, combinations of information technology and education being commonly used in teaching [1-2]. Furthermore, the introduction of information technology can enhance adaptive learning, since it allows students to have more learning opportunities [3-4]. In addition, Tseng [21] indicated that online tests are more convenient than traditional paper and pencil tests (P&P tests). Many studies have confirmed that the combination of appropriate digital learning tools and materials can promote learning motivation in students [23-25] and enhance their learning performance in English courses [26]. Through the convenience of online platforms, teachers can effectively evaluate and monitor students' academic performance virtually [27].

Although online tests can bring learning benefits to English courses, they may also cause some problems related to English learning. One of the issues is whether the level of the test items in online tests is consistent with the learner's ability level. To better implement adaptive online testing, computerized adaptive testing (CAT) has been developed to resolve issues where traditional computerized tests use inappropriate test items. In addition, CAT makes full use of computer operations, storage, and transmission to dynamically select test items that meet the current ability of the examinee and provides test content specific to the individual [5-6].

One of the important conditions for CAT to function properly is estimating the difficulty of the test items correctly, and the item response theory (IRT) is usually used as the mathematical model to estimate the item difficulty index [7]. It is not difficult to achieve successful computerized adaptive testing if each test item has already been given its corresponding difficulty level. However, it is suggested here that determining how to quickly adjust the difficulty of individual test items among a large number of test items has become a very difficult problem for CAT to solve. In a test system with a large number of test items, it may be difficult to obtain the number of pre-test samples required by the IRT model for parameter estimation of the test items, and when increasing the test items, it is necessary to re-estimate the parameters of the test items, which results in the generation of big data that is difficult to analyze manually. To solve this problem, it has become a trend to automatically analyze big data [8]. Big data are currently an important research field contributing to the development

*Corresponding Author: Yueh-Min Huang; E-mail: huang@mail.ncku.edu.tw
DOI: 10.53106/160792642021122207013

of internet and web technologies [9].

However, it was found here that the difficulty of English grammar items could not be automatically adjusted in previous studies, as well as the fact that the test items in big data collected by online learning platforms creates the biggest challenge related to adjusting item difficulty. To solve the limitations of previous research, in this study, an adaptive English learning platform is developed, and a method to automatically estimate an item difficulty index based on IRT is proposed. A big data analysis of a large number of test question answer records collected by students on an adaptive English learning platform is used to automatically adjust the difficulty index of each test item.

Section 2 reviews and discusses the relevant theoretical background; Section 3 introduces the research methods of this study, including the adaptive English learning platform developed by this study, the formulation of test difficulty, and the experimental methods; Section 4 describes the data analysis and interpretation of the experimental results of this study; Section 5 discusses and summarizes this study.

2 Literature Review

This section reviews and discusses the basic theory and related applications of the test difficulty, computerized adaptive testing, and the item response theory. According to the survey of this study, no study can automatically adjust the difficulty of test items according to the students' answer records and provide appropriate questions according to the students' ability values to carry out computerized adaptive testing. The following is a detailed overview of the application of various theories and related research.

2.1 Item Response Theory

The item response theory (IRT) evaluates the ability of the examinee or the position of a continuous range of psychological dimensions, as based on the information reflected by the test items, and is a psychometric theory that describes the position of the examinee's ability level in the scale space according to the examinee's response to individual test items. In addition, IRT uses item information to express the accuracy of the test. The higher the amount of information, the more accurate the test is to measure the ability position. According to the local independence assumption of IRT, the sum of all the item information of an examinee on the test paper is test information. This means that, when testing, it is not necessary to provide the same test items to all the examinees, meaning it can be used flexibly according to individual needs, which makes the test more efficient. This is the basic concept of CAT according to the actual situation. The larger the amount of the information provided by the test to the examinee, the smaller the measurement error of the test to the examinee, and the more accurate the ability position estimation [10].

On the other hand, the commonly used estimation methods of the ability parameters of the examinees are Maximum likelihood estimation (MLE) method [29] and Maximum a posteriori Bayesian (MAPB) estimation [30]. The commonly used estimation methods of the test items are Joint maximum likelihood (JML) estimation [31], Marginal maximum likelihood estimation [32], and Conditional maximum likelihood estimation [33].

This study used the model proposed by Rasch [11]. This model is expressed as $p(X=1|\theta,b)=\frac{e^{\theta-b}}{1+e^{\theta-b}}$ Where X is the score of the subjects to answer an item. For example, X=1 means the subject answers the item response correctly, X=0 means the subject did not answer the item response correctly; θ is the ability value of the subjects; b is the difficulty parameter of an item.

2.2 Item Difficulty Index

There are two commonly used methods to estimate the item difficulty index. First, the item difficulty index can be expressed by the percentage of all the examinees who answered correctly or passed the test [12-13]. Second, the examinees can be ranked according to the total score of the test, then the two groups with the highest score and the lowest score can be designated as the high score group and low score group, and then, the percentage of the two groups who answered correctly or passed a test can be calculated, respectively. Finally, the average percentage of the two groups can be taken as the item difficulty index. The higher the P-value, the lower the degree of difficulty; the smaller the P-value, the higher the degree of difficulty [14-15].

2.3 Computerized Adaptive Testing

The traditional test method is to test all examinees with the same set of test items; however, the same items are not appropriate for some types of tests, meaning it may be too difficult or too simple for examinees with high or low abilities to test with the same items. As it is impossible to accurately identify the ability level of the examinees by testing with inappropriate items, the significance of the test is lost. In order to improve the weakness of traditional testing, the basic concept of Computerized Adaptive Testing (CAT) is to select the test items that are most suitable for the current ability level of the examinees for testing. Whenever an item is completed, the test system will immediately evaluate the ability level of the examinee, and take this evaluation as the basis for selecting the next test item, that is to say, whether the examinee

answers correctly or not will affect the difficulty of the next test item. As CAT is a test method specially designed for individuals, this dynamic selection strategy can be used according to the ability level of different examinees. Due to this dynamic selection strategy, which determines the difficulty of test items based on the ability level of the examinees, CAT can shorten the test length and accurately evaluate the ability level of the examinees, in order to achieve the goal of testing according to individual ability [16-17].

Triantafillou, et al. [18] integrated the function of computerized adaptive testing into mobile devices by means of mobile technology to develop the CAT-MD learning tool and explored the process of students' use of CAT-MD in physical subjects. The study showed that CAT-MD can provide accurate results according to the difficulty of the test items. In addition, the advantage of using a mobile device is that it can be operated anywhere. Čisar, et al. [19] used computer adaptive tests to evaluate students' knowledge of program courses and compared the differences between paper and pencil tests (P&P tests) and computer adaptive tests. The research results showed that students who use computer adaptive tests can get higher scores. Compared with P&P tests, these students are more able to maintain a good attitude in class, and they experience less pressure in the computerized adaptive test. It was pointed out that information technology can provide more learning environments for a second language and that computer adaptive testing can adjust the range of questions according to the survey of students [20]. Tseng [21] measured the feasibility of evaluating the level of English vocabulary knowledge through computerized adaptive testing, and explored the differences between CAT and P&P tests. The study showed that CAT can not only replace traditional P&P tests but also perform better than P&P tests in vocabulary estimation.

According to the previous discussion, most studies have confirmed the effectiveness of CAT. However, in studies focused on English education, there has been no discussion of the application of grammar test items. Also, determining how to adjust the difficulty of test items based on student answers has not been deeply discussed in studies on language learning, and determining how to automatically adjust the difficulty of a large number of test items is a challenge in computerized adaptive testing. This study thus provides an exploration of how to analyze and automatically adjust item difficulty based on a large amount of answer records, so as to solve the problem of dynamic adjustment of test difficulty. In addition, this study is intended to verify whether the ability value automatically assessed by the CAT is the same as the ability value obtained with the P&P tests in order to calculate the accuracy of the system. Finally, this study also provides an exploration of students' English learning performance.

The research questions addressed in this study are as follows:

1. Through analyzing the answer records of the test items through big data, is it feasible for the automatic test item difficulty index estimation method to be used in the CAT?
2. What is the degree of conformity between the ability value automatically assessed by the CAT and the ability value obtained using the P&P test?
3. Do students exhibit improved English learning performance after they use the adaptive English learning platform to conduct the CAT?

3 Method

This study developed an adaptive English learning platform, in which CAT is based on IRT to design a method for estimating an item difficulty index. By analyzing students' test records, the system can adjust the difficulty index of each test in an automatic manner, and then, automatically provide the appropriate test difficulty according to students' ability values to achieve the purpose of CAT. Moreover, according to the automatic adjustment method of the difficulty of test items, this study can effectively analyze the answer records of big data, spare the evaluation time of experts and teachers, and expand the test item database at any time to automatically estimate and quickly define the difficulty of new test items.

3.1 Automatically Adjusting Item Difficulty Index

In this study, 53,191 test answering records were collected as the basis of the estimation method for the item difficulty index, and the item difficulty index could be divided into N levels. In addition, the ability evaluation of the examinee was also considered in the estimation process of the item difficulty index. Therefore, before analyzing the answering records, this study gives different expectation values according to the model proposed by Rasch [11] for the right answer probability of the following different ability values corresponding to different test levels. The Eq. (1) is as follows:

$$p_i(\theta) = \frac{e^{\theta - b_i}}{1 + e^{\theta - b_i}} \quad \textbf{(1)}$$

The examinee of different ability values and the difficulty index of a certain test item is expressed by the correct answer rate, and the calculation is expressed by Eq. (2).

$$Diff_{il}^{k} = \left| \frac{R_{il}^{k}}{N_{il}^{k}} - p_i(\theta) \right| \quad \textbf{(2)}$$

where $Diff_{il}^{k}$ is the abnormal rate of correct answers of

group k whose difficulty index is level l in item i, R_{il}^{k} is the number of the correct answers in group k with difficulty level l for item i, and N_{il}^{k} is the total number of group k with difficulty index level l for item i.

The sum of $Diff_{il}^{k}$ is the abnormal rate of correct answers of group k when the difficulty index is level l in item i, and its calculation is expressed in Eq. (3).

$$Diff_{il} = \sum_{k=1}^{n} Diff_{il}\, k \tag{3}$$

where $Diff_{il}$ is sum of the abnormal rate of correct answers when the difficulty index is level l. When the abnormal rate of correct answers of the difficulty index of a test item is the smallest, it is the item difficulty index of the test item, which is calculated by Eq. (4). Where, D is the difficulty index of a test item.

$$D = \min(Diff_{i1}, Diff_{i2}, ..., Diff_{in}) \tag{4}$$

3.2 Adaptive English Learning Platform

An adaptive English learning platform provides 1,517 English test items for students to engage in CAT. Each test record is stored in the database by the platform, including the number of correct answers, number of test items, personal ability values, test time, etc. In addition, the number of English searches and vocabulary practice are recorded. This study divided the test items and personal ability values into 9 levels, ranging from 0.1 to 0.9. Table 1 is an example of the test items in the database. These data include the difficulty level of test items, number of correct answers, number of incorrect answers, and the total answers.

Table 1. Example of the test items in the database

item #	difficulty level of test items	number of correct answers	number of incorrect answers	total answers	number of adjustments
1	0.1	286	246	532	7
2	0.2	276	354	630	5
3	0.4	262	275	537	5
4	0.8	200	273	473	8
5	0.9	216	246	462	6
...	...	...	...	...	...

This platform is based on the IRT to design an estimation method of the automatized item difficulty index in order to achieve CAT. It provides grammar and vocabulary tests, which comprise a total of 1,517 English test items. The grammar tests and vocabulary tests are based on student ability values, meaning the system automatically adjusts the difficulty index of the test bank according to the students' answer records to realize the purpose of CAT.

3.3 Participants and Experiment Procedure

This study investigated 162 undergraduate students in the department of computer science and information engineering at a university in Taiwan. The researcher conducted experiments and data collection in a professional English course. The participants were 20 years old. The researcher and teacher coordinated the experiments and the course. All students have a background in information engineering and participated in the experiment voluntarily. In addition, before the experiment, none of the participants had engaged in CAT activities previously in a professional English course.

The experimental procedure is shown in Figure 1. The participants took a 20-minute pre-test to test their prior knowledge of English. Every week, the teacher invited a total of 162 students to take a 30-minute CAT on the adaptive English learning platform. The CAT includes both a grammar test and a vocabulary test. All of the participants used the adaptive English learning platform for a 16-week CAT in the professional English course. After 16 weeks of adaptive English learning on the platform for professional English learning, the participants took a 20-minute post-test to complete this experimental activity.

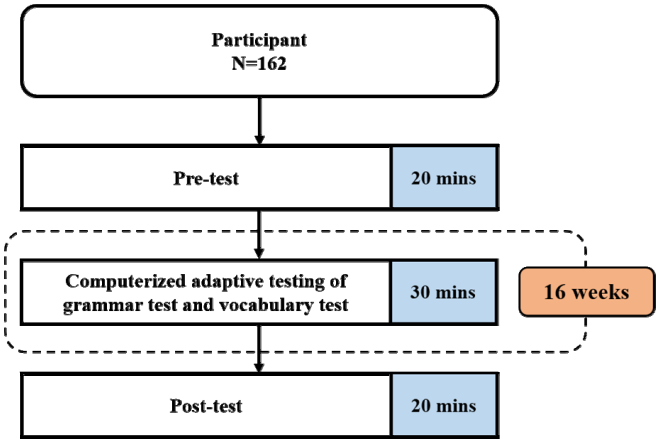

Figure 1. Experimental procedure of this study

3.4 Data Collection and Data Analysis

• Data collection

The pre-test and post-test results for 162 students were collected, and the scores were normalized. The normalized scores ranged from 0 to 1. In addition, 1,517 English test items on the adaptive English

learning platform were provided, and 53,191 item answer records were collected from the 162 students who took the CAT on this platform.

• Data analysis

According to the research questions, four different data analyses were conducted, which are described as follows:

1. The adaptive English learning platform developed in this study provided 1,517 English test items, and 53,191 item answer records that were collected through the CAT were used to calculate and analyze the number of automatically adjusted difficulty levels for all test items and the average number of adjustment convergences.

2. The students' ability values automatically assessed in CAT were collected as well as the P&P test scores for the post-test. Then, the students' ability values and P&P test scores were divided into high ability and low ability groups based on a percentile rank of 50. Through this method, the values for the high-ability groups (the students' ability value automatically assessed by the CAT and the students' ability values in the P&P test) and the low-ability groups (the students' ability values automatically assessed by the CAT and the students' ability values for the P&P test) could be used to verify whether the ability values obtained through automatic assessment in CAT were the same as those obtained using the P&P test and determine the system accuracy of the CAT.

3. A paired sample t-test was used to analyze the pre-test and post-test P&P test scores of 162 students in order to explore the learning performance of students after the use of CAT on the adaptive English learning platform.

4 Results

The adaptive English learning platform of this study can automatically determine the difficulty level of test items based on item response theory to effectively realize the CAT. In order to evaluate the feasibility of automatically adjusting the difficulty levels of test items, and to explore the learning performance of an adaptive English learning platform combined with CAT for students in the professional English course. This study collected 162 students' usage behaviors of the adaptive English learning platform, analyzed 53,191 answer records, discussed the adjustment times and convergence of the difficulty level of the test items. In order to verify the degree of conformity between the student's ability value automatically assessed by the CAT and the student's ability values of the P&P test, this study evaluated the accuracy of CAT of the adaptive English learning platform. In addition, this study analyzes the pre-test and post-test scores of all students to evaluate whether students can improve their learning performance through using the adaptive English learning platform.

4.1 Automatically Adjusting the Item Difficulty Index

The first research question addressed in this study was to explore the feasibility of CAT in automatically adjusting the difficulty levels of test items. Therefore, this study was based on the use of the IRT to automatically adjust the difficulty levels of the test items. The procedure used to automatically adjust the difficulty levels of the test items is shown in Figure 2. On the adaptive English learning platform developed in this study, an initial value for the difficulty of the 1,517 test items in the database is set first. After the students had undergone CAT, the adaptive English learning platform automatically analyzed all English test items and the 53,191 item answer records collected through CAT. Using the automated test item difficulty index estimation method, this platform could automatically calculate each English test item and continuously adjust the difficulty level to the final test item. After automatic calculation and adjustment, the final difficulty level of each test item was automatically updated in the database.

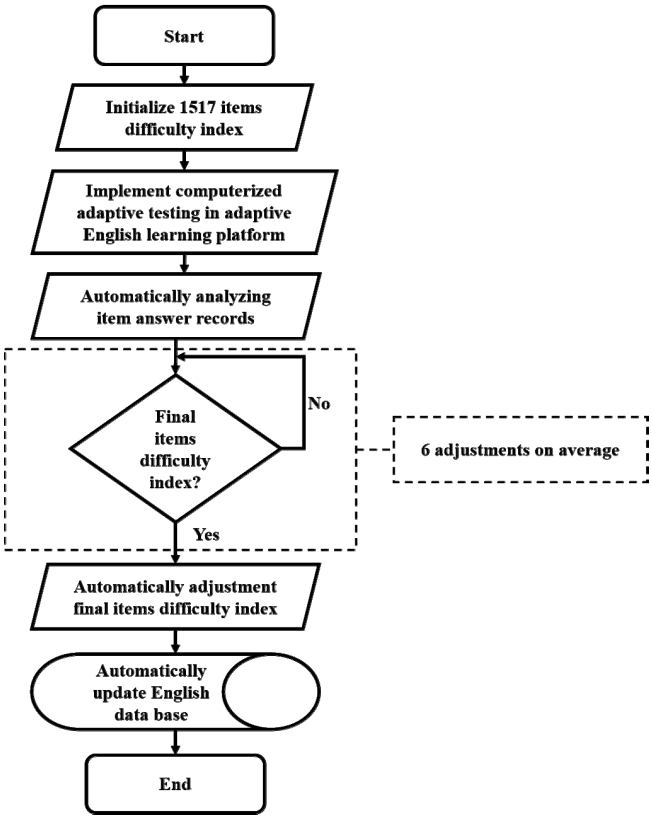

Figure 2. Flowchart of automatically adjusting difficulty levels of test items

A total of 1,517 English test items in the database were analyzed, and the item difficulty index was divided into 9 levels. Based on 53,191 item answer records, the system automatically adjusted the difficulty levels of the test items a total of 9,385 times.

In addition, when an item difficulty level had been adjusted six times, on average, the final difficulty level of the test item was acquired. In other words, when an examinee conducted a test of new test items on the platform, each test item had to be adjusted about six times before it would converge to the final difficulty level of the test item. In addition, the system proposed by this study will automatically adjust the difficulty of the test items every week (as shown in Figure 3).

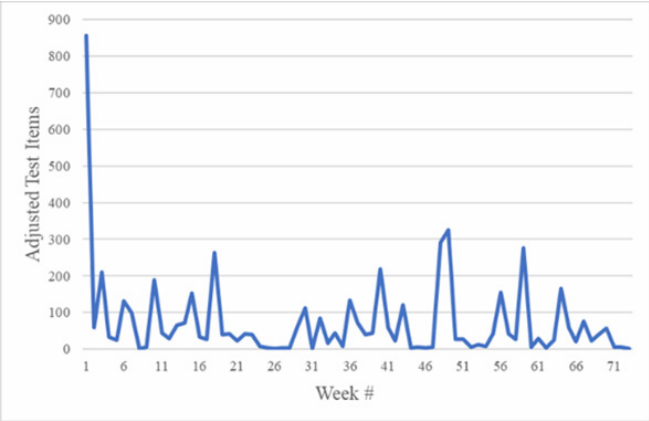

Figure 3. The trend of the difficulty adjustment

For the adaptive English learning platform developed in this study, CAT can provide appropriate test items according to the students' ability values and can record the students' answers. According to the students' answer records, the system automatically adjusts the difficulty levels of the test items according to the IRT and updates the setting values of the difficulty level of the test item in the database. Based on the IRT, the more item information there is, the higher the accuracy of the difficulty assigned by the system to the examinee will be. Based on this result, the system saves expert evaluation time by automatically determining and adjusting the difficulty levels of the test items and quickly achieving a stable weighting through the use of a large amount of answer data, thus allowing the difficulty levels of the test items to meet the students' ability levels more quickly.

4.2 Accuracy of Computerized Adaptive Testing in Adaptive English Learning Platform

The second research question addressed in this study was an effort to determine the degree of conformity between the ability values automatically assessed by CAT and the ability values obtained from the results of the P&P test. Therefore, ability values for 162 students automatically assessed using CAT were collected, ranked from high to low, and classified into high ability and low ability groups with a Percentile Rank of 50%. The scores on the P&P tests for the 162 students were also sorted from high to low and divided into high ability and low ability groups according to a Percentile Rank of 50% in order to use the ability values of the students to calculate and verify the accuracy of the CAT.

In this study, the ability values of the students automatically assessed with CAT were compared with their ability values from the P&P tests. If the ability value of an examinee automatically assessed by CAT and the P&P test scores in the post-test were in the group with the same level of ability, this indicated that the proposed system could accurately and automatically determine the ability values of the examinee and provide the appropriate difficulty levels for the test items, which was also in line with the purpose of CAT. As shown in Table 2, after the comparison made in this study, 140 students' ability values automatically assessed using CAT were the same as those obtained with the P&P test, so the accuracy of the platform was as high as 86%, indicating that the accuracy of CAT provided by the adaptive English learning platform was as high as 86%. In addition, there were a total of 82 students from the high ability group, of which 74 students' ability values automatically assessed by CAT were consistent with the ability values obtained using the P&P test, for an accuracy rate of 90%. There were a total of 80 students from the low ability group, of which 66 students' ability values automatically assessed using CAT were consistent with the ability values obtained with the P&P test, giving an accuracy rate of 82%.

Table 2. The accuracy of the student ability automatically assessed with CAT and that obtained with the P&P test

	number of the students automatically assessed by CAT	number of the students	accuracy
low ability group	66	80	82% (60/80)
high ability group	74	82	90% (74/82)
total	140	162	86% (140/162)

4.3 English Learning Performance

The third research question of this study was intended to explore whether students could improve their English learning performance by using the CAT proposed for use on this platform. Therefore, this study the collected pre-test and post-test P&P test scores for the 162 students were normalized in a range from 0 to 1, and a paired sample t-test analysis was conducted. Table 3 shows the results of the paired sample statistics. The mean of the pre-test was 0.52, and the standard deviation was 0.22. The mean of the post-test was 0.61,

and the standard deviation was 0.22.

Table 3. Paired sample statistics of pre-test and post-test

	N	*M*	*SD*
Pre-test	162	0.52	0.22
Post-test	162	0.61	0.22

According to the results shown in Table 2, the mean of the post-test was higher than the mean of the pre-test. Table 4 shows that the difference between the mean of the post-test and pre-test reached a significant difference (t = -7.057, $p < 0.001$), which means that the CAT on the adaptive English learning platform indeed enhanced the students' English learning performance.

Table 4. Paired sample *t*-test result of pre-test and post-test

	M	*SD*	*df*	*t*
Pre-test - Post-test	-0.087	0.012	161	-7.057***

*$p < .05$, **$p < .01$, ***$p < .001$.

5 Discussions and Conclusions

In this study, an adaptive English learning platform and a method to automatically estimate the item difficulty index based on the IRT were developed. The purpose was to save expert evaluation time by automatically determining and adjusting the test item difficulty through the analysis of big data, so as to achieve the purpose of the use of CAT more efficiently and also to effectively improve students' English learning performance.

Despite the fact that previous researchers were unable to evaluate the difficulty of test items in an automated way, in the current study, the difficulty level of 1,517 English test items was automatically calculated, where the difficulty level was adjusted 9,385 times in order to successfully obtain the best difficulty levels for the test items after an average of six adjustments based on 53,191 item answer records. From the results of this study, it was apparent that the database achieved a convergence effect with six updates, on average. According to the first research question, the results of this study confirmed the actual benefits and feasibility of automatically adjusting the difficulty levels of the test items, meaning it is not necessary to manually evaluate tens of thousands of answer records. In addition, 53,191 item answer records of the test items were collected through the CAT. By automatically adjusting the difficulty of the test items, it was easy to use the huge amount of answer records to quickly evaluate the best difficulty of the test items, so the students' ability values could be automatically assessed through CAT and items could be provided that were consistent with their ability. According to the automatic item difficulty index estimation method, CAT on the adaptive English learning platform made it possible to automatically adjust the difficulty index of each test item and automatically calculate the difficulty index of each test item and the average number of convergences through the system. This method effectively reduced the time required for manually estimating the difficulty levels of the test items as well as the limit of the number of test items. In the future, with increases in new item banks or through student usage, test difficulty could be re-estimated by automatically adjusting it, thereby effectively improving efficiency and reducing the burden and errors associated with manual evaluations. This result confirms the proposed system's contribution to the development of technology and to effectively moving the future development of adaptive systems in English education forward based on previous research [22].

The second research question of this study was intended to explore the degree of conformity between the ability values automatically assessed by CAT and the ability values obtained from the results of the P&P test. According to the results, among the 162 students, 140 had the same ability value automatically assessed by both CAT and the P&P test. This means that CAT had a system accuracy of 86% (among which, the accuracy of the high ability group was 90%, and the accuracy of the low ability group was 82%). In other words, the higher accuracy of the system calculated by this method means that the ability values, as automatically estimated by the CAT, are similar to those obtained using the P&P test. This also means that this system can provide tests at different difficulties for different students, thus, effectively achieving the purpose of CAT and confirming the effectiveness of the CAT in learning assessments, which is consistent with the viewpoints mentioned in previous studies [5, 21].

As stated above, the CAT used on the adaptive English learning platform not only automatically adjusts the difficulty levels of the test items, but also automatically assesses student ability so as to provide test items that match the ability levels of students and achieve the ultimate purpose of CAT. Therefore, in the present study the pre-test and post-test scores of 162 students were analyzed to explore whether their English learning performance could be improved by students using CAT on the adaptive English learning platform. According to the results and addressing the third research question posed in this study, students using an adaptive English learning platform for CAT can significantly improve their English learning performance in professional English courses. This is consistent with the viewpoints mentioned in previous studies [28]. It was also found from the evaluation of English learning performance that students using CAT in the adaptive English learning platform significantly

improve their post-test scores and learning progress.

In summary, the IRT, and big data technology were used in the present study to automatically adjust the difficulty levels of the test items and resolve the limitations characteristic of previous research. A method was proposed to estimate and automatically adjust the item difficulty index based on the IRT, which not only resolved the challenge of manual evaluation brought by big data but also automatically and efficiently determined the test difficulty in items in a large database. An objective analysis was made of answered items based on big data records in order to automatically adjust the item difficulty index in all test item databases. This reduces that time required for experts and teachers to quickly determine the difficulty of test items as well as the time required for manual adjustment of item difficulty to bring the items more in line with the ability levels of students and achieve the purpose of CAT. These students can also improve their English learning performance through the use of CAT.

References

[1] Z. I. Abrams, Collaborative writing and text quality in Google Docs, *Language Learning & Technology*, Vol. 23, No. 2, pp. 22-42, June, 2019.

[2] M.-P. Chen, L.-C. Wang, D. Zou, S.-Y. Lin, H. Xie, Effects of caption and gender on junior high students' EFL learning from iMap-enhanced contextualized learning, *Computers & Education*, Vol. 140, Article No. 103602, October, 2019.

[3] K. Pliakos, S.-H. Joo, J. Y. Park, F. Cornillie, C. Vens, W. Van den Noortgate, Integrating machine learning into item response theory for addressing the cold start problem in adaptive learning systems, *Computers & Education*, Vol. 137, pp. 91-103, August, 2019.

[4] H. M. Truong, Integrating learning styles and adaptive e-learning system: Current developments, problems and opportunities, *Computers in human behavior*, Vol. 55, pp. 1185-1193, February, 2016.

[5] H.-H. Chang, Psychometrics behind computerized adaptive testing, *Psychometrika*, Vol. 80, No. 1, pp. 1-20, March, 2015.

[6] H. Wainer, G. L. Kiely, Item clusters and computerized adaptive testing: A case for testlets, *Journal of Educational measurement*, Vol. 24, No. 3, pp. 185-201, September, 1987.

[7] W. J. Van der Linden, P. J. Pashley, Item selection and ability estimation in adaptive testing, in: W. Van der Linden, C. Glas (Eds.), *Elements of adaptive testing*, Springer, 2009, pp. 3-30.

[8] C.-W. Tsai, C.-F. Lai, H.-C. Chao, A. V. Vasilakos, Big data analytics: a survey, *Journal of Big data*, Vol. 2, No. 1, pp. 1-32, October, 2015.

[9] N. A. Ghani, S. Hamid, I. A. T. Hashem, E. Ahmed, Social media big data analytics: A survey, *Computers in Human Behavior*, Vol. 101, pp. 417-428, December, 2019.

[10] F. M. Lord, Applications of item response theory to practical testing problems, Routledge, 2012.

[11] G. Rasch, Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests, Nielson and Lydiche, Copenhagen, 1960.

[12] H. M. Abdulghani, F. Ahmad, G. G. Ponnamperuma, M. S. Khalil, A. Aldrees, The relationship between non-functioning distractors and item difficulty of multiple choice questions: A descriptive analysis, *Journal of Health Specialties*, Vol. 2, No. 4, pp. 148-151, October, 2014.

[13] M. R. Hingorjo, F. Jaleel, Analysis of one-best MCQs: the difficulty index, discrimination index and distractor efficiency, *JPMA-Journal of the Pakistan Medical Association*, Vol. 62, No. 2, pp. 142-147, February, 2012.

[14] T. M. Haladyna, Developing and validating multiple-choice test items, Routledge, 2004.

[15] H. K. Suen, *Principles of test theories*. Routledge, 2012.

[16] S.-C. Cheng, Y.-T. Lin, Y.-M. Huang, Dynamic question generation system for web-based testing using particle swarm optimization, *Expert systems with applications*, Vol. 36, No. 1, pp. 616-624, January, 2009.

[17] Y.-M. Huang, Y.-T. Lin, S.-C. Cheng, An adaptive testing system for supporting versatile educational assessment, *Computers & Education*, Vol. 52, No. 1, pp. 53-67, January, 2009.

[18] E. Triantafillou, E. Georgiadou, A. A. Economides, The design and evaluation of a computerized adaptive test on mobile devices, *Computers & Education*, Vol. 50, No. 4, pp. 1319-1330, May, 2008.

[19] S. M. Čisar, P. Čisar, R. Pinter, Evaluation of knowledge in Object Oriented Programming course with computer adaptive tests, *Computers & education*, Vol. 92-93, pp. 142-160, January-February, 2016.

[20] S. Bodnar, C. Cucchiarini, H. Strik, R. van Hout, Evaluating the motivational impact of CALL systems: current practices and future directions, *Computer Assisted Language Learning*, Vol. 29, No. 1, pp. 186-212, 2016.

[21] W. T. Tseng, Measuring English vocabulary size via computerized adaptive testing, *Computers & Education*, Vol. 97, pp. 69-85, June, 2016.

[22] V. Slavuj, A. Meštrović, and B. Kovačić, Adaptivity in educational systems for language learning: a review. *Computer Assisted Language Learning*, Vol. 30, No.1-2, pp. 64-90, 2017.

[23] C. M. Chen, L. C. Chen, S.M. Yang, An English vocabulary learning app with self-regulated learning mechanism to improve learning performance and motivation, *Computer Assisted Language Learning*, Vol. 32, No. 3, pp. 237-260, 2019.

[24] C. S. Huang, S. J. Yang, T. H. Chiang, A. Y. Su, Effects of situated mobile learning approach on learning motivation and performance of EFL students, *Journal of Educational Technology & Society*, Vol. 19, No. 1, pp. 263-276, January, 2016.

[25] C. H. Su, C. H. Cheng, A mobile gamification learning system for improving the learning motivation and achievements, *Journal of Computer Assisted Learning*, Vol. 31, No. 3, pp. 268-286, June, 2015.

[26] Y. C. Kuo, H. C. Chu, M. C. Tsai, Effects of an integrated

physiological signal-based attention-promoting and English listening system on students' learning performance and behavioral patterns, *Computers in Human Behavior*, Vol. 75, pp. 218-227, October, 2017.

[27] H. K. Wu, C. Y. Kuo, T. H. Jen, Y. S. Hsu, What makes an item more difficult? Effects of modality and type of visual information in a computer-based assessment of scientific inquiry abilities, *Computers & Education*, Vol. 85, pp. 35-48, July, 2015.

[28] M. Sharifi, A. R. AbuSaeedi, M. Jafarigohar, B. Zandi, Retrospect and prospect of computer assisted English language learning: a meta-analysis of the empirical literature, *Computer Assisted Language Learning*, Vol. 31, No. 4, pp. 413-436, 2018.

[29] I. J. Myung, Tutorial on maximum likelihood estimation, *Journal of Mathematical Psychology*, Vol. 47, No. 1, pp. 90-100, February, 2003.

[30] A. F. van der Meer, M. A. Marcus, D. J. Touw, J. H. Proost, C. Neef, Optimal sampling strategy development methodology using maximum a posteriori Bayesian estimation, *Therapeutic drug monitoring*, Vol. 33, No. 2, pp. 133-146, April, 2011.

[31] Y. Chen, X. Li, S. Zhang, Joint maximum likelihood estimation for high-dimensional exploratory item factor analysis, *Psychometrika*, Vol. 84, No. 1, pp. 124-146, March, 2019.

[32] R. D. Bock, M. Aitkin, Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm, *Psychometrika*, Vol. 46, No. 4, pp. 443-459, December, 1981.

[33] E. B. Andersen, Asymptotic properties of conditional maximum-likelihood estimators, *Journal of the Royal Statistical Society: Series B (Methodological)*, Vol. 32, No. 2, pp. 283-301, July, 1970.

Biographies

Shu-Chen Cheng received the B.S. and Ph.D. degrees in Engineering Science from National Cheng-Kung University, Taiwan, in 1989 and 2004. She is a Professor at Department of Computer Science and Information Engineering in Southern Taiwan University of Science and Technology, Taiwan. Her research interests include artificial intelligence, data mining, text mining, and image processing.

Yu-Ping Cheng received the B.S. degree in Department of Computer Science and Information Engineering in Southern Taiwan University of Science and Technology, Taiwan, in 2016, and received the Ph.D. degree in Engineering Science from National Cheng-Kung University, Taiwan, in 2021. His research interests include e-learning, artificial intelligence, and data mining.

Yueh-Min Huang received the M.S. and Ph.D. degrees in Electrical Engineering from the University of Arizona in 1988 and 1991 respectively. He is a Chair Professor in Department of Engineering Science, National Cheng-Kung University, Taiwan. He has co-authored 3 books and has published more than 300 refereed journal research papers. His research interests include e-Learning, multimedia communications, wireless networks, and artificial intelligence. He is in the editorial board of several international journals in the area of educational technology, computer communications, and web intelligence. He was also serving as the directors of Disciplines of Applied Science Education and Innovative Engineering Education in Taiwan's Ministry of Science and Technology. Dr. Huang is a senior member of the IEEE and became Fellow of British Computer Society in 2011. Dr. Huang has received many research awards, such as Taiwan's National Outstanding Research Award in 2011 and 2014, given to Taiwan's top 100 scholars. Dr. Huang is also the Honorary Chair of the International Conference of Innovative Technologies and Learning (ICITL). According to a paper published in BJET, he is ranked no. 3 in the world on terms of the number of educational technology papers published in the period 2012 to 2017.