A CEFR-BASED COMPUTERIZED ADAPTIVE TESTING SYSTEM FOR CHINESE PROFICIENCY

Hsuan-Po Wang
Affiliation: Graduate Institute of Educational Measurement and Statistic
University: National Taichung University of Education, Taiwan
E-mail: sam710701@gmail.com

Prof. Bor-Chen Kuo
Affiliation: Graduate Institute of Educational Measurement and Statistic
University: National Taichung University of Education, Taiwan
E-mail: kbc@mail.ntcu.edu.tw

Prof. Ya-Hsun Tsai
Affiliation: College of International Studies and Education for Overseas Chinese
University: National Taiwan Normal University, Taiwan
E-mail: yahsun@ntnu.edu.tw

Prof. Chen-Huei Liao
Affiliation: Department of Special Education
University: National Taichung University of Education, Taiwan
E-mail: chenhueiliao@gmail.com

ABSTRACT

In the era of globalization, the trend towards learning Chinese as a foreign language (CFL) has become increasingly popular worldwide. The increasing demand in learning CFL has raised the profile of the Chinese proficiency test (CPT). This study will analyze in depth the inadequacy of current CPT's utilizing the common European framework of reference (CEFR) for language learning, teaching, and assessment to develop a set of reliability and validity standards for a computerized adaptive testing (CAT) CPT system. Actual performance of computerized tests will simulate the empirical data via the CAT system process and assess the efficacy of this system.

Keywords: Chinese Proficiency Test, Common European Framework of Reference, Computerized Adaptive Testing

INTRODUCTION

With the growing demand of learning Chinese as a foreign language (CFL), the development and utility of the proficiency test for "non-native Chinese" learners is essential, particularly, countries that are in the preliminary stage of promoting CFL courses in the educational institutions and organizations. For example, United Kingdom language school has included CFL in its foreign language learning curriculum. National Security Language Initiative (NSLI) of the United States has identified Chinese language an important national security strategic language, and is planning on including Chinese in the foreign language learning curriculum in schools and workplaces (U.S. Department of State, 2006). All these programs show that learning CFL is becoming an important issue due to the large demand of Chinese language proficiency. Currently, Test of Chinese as a Foreign Language (TOCFL), Hanyu Shuiping Kaoshi (HSK), Test of Practical Chinese (C. Test), Scholastic Assessment Test (SAT) subject test in Chinese with listening, and Advanced Placement (AP) Chinese language and culture are often used to assess Chinese proficiency (SC-TOP, 2011; HSK, 2011; C. Test, 2011; College Board, 2011a; College Board, 2011b). However, the majority of these tests are administered by the traditional paper and pencil tests (PPT) format. Although there are many studies about developing the tools for learning CFL (Wong, Gao, Chai, & Chin, 2011; Zhao, Wang, Wu, & He, 2011; Shieh, 2011), the construction of computerized test for CFL is hard to find on the literatures. The aims of the present study are: adopting the Common European Framework of Reference (CEFR) for item development; providing a framework by using item response theory (IRT) as the scoring method; constructing computerized adaptive testing (CAT) system.

Currently the frameworks of reference are inconsistent among CPTs which results in various classification levels of proficiency and makes it difficult to classify learners' language levels consistently. With a international recognized common reference framework that describe learner language skills in detail, it will make it easier to identify learner's status of Chinese learning. In addition, this framework will allow learners to continue their Chinese language learning in any countries without extraneous evaluation of level assessment. By providing a common basis for the explicit description of objectives, content and methods, the CEFR will enhance the transparency of courses, syllabuses and qualifications, thus promoting international co-operation in the field of modern languages. The provision of objective criteria for describing language proficiency will facilitate the mutual recognition of qualifications in different learning contexts and settings, and accordingly will promote the mobility of learning CFL. Currently, proficiency tests that adopt CEFR as the framework reference are: Test of English for International Communication (TOEIC), Test of English as a Foreign Language (TOEFL), Cambridge Main Suite, Business Language Testing Service (BULATS), Test Deutsch als Fremdsprache (Test Daf), Japanese-Language Proficiency Test (JLPT), DiplômeD'Etudes en Langue Française (DELF), and HSK etc. (Kecker & Eckes, 2007; Tannenbaum & Wylie, 2005).

The majority of these CPTs analyze examinee's proficiency level based on the classical test theory (CTT), which use the total of observed scores (raw score) to classify examinees CFL proficiency scales (SC-TOP, 2011; HSK, 2011; C. Test, 2011; NCACLS, 2010). Using raw score to represent proficiency scale of test violates the following assumptions: meaningful measurement, unidimensionality, and linearity of data characteristics (Lord, 1980; Wright, 1999). In addition, when an examinee participates in different proficiency tests, the total raw score of each test will not be able to reflect truthfully his/her language skills since the difficulty of test items are undefined and uninformed. However, the psychometric basis of tests has changed dramatically, even though CTT has been used for several decades, the application of IRT increases rapidly and become the mainstream of measurement theory. Recently, more standardized tests are developed by using IRT because of its theoretical measurement principles. IRT consists of mathematical models designed to describe the performance of examinees on test items. IRT not only selects the most appropriate items for examinees, it also equates scores across different subsets of items. For example, the Scholastic Assessment Test (SAT) and the Graduate Record Examination (GRE) both use IRT for ability estimation.

Using computers to deliver standards-based assessments is becoming common among education departments, legislators, and policy makers. Computer-based testing (CBT) has become one of most common forms of testing since 1990s. CBT has been developing quickly since then as new question formats, alternative models of measurement, improvements in test administration, immediate feedback to test takers, more efficient information gathering (Akdemir & Oguz, 2008; Mills, 2002; Wise & Plake, 1990), and development of new methods of assessment such as simple adaptations of multiple-choice items to more innovative item types (Jodoin, 2003). Through the use of multimedia technology, CBT is able to apply more diverse and developed items which are closer to real situations. CAT combines the multimedia characters of computerized testing and the efficiency of adaptive testing. With computerization, color, sound, animation, interaction, and performance could be integrated in a test. This will definitely improve the validity of the test. Thus, although the majority of these CPTs are administered by the traditional PPT, College Board is conducting an important project in developing CBT (College Board, 2011a). Although there are CPTs, which were developed based on CBT, yet, CAT was not used. With CAT, precision of abilities can be obtained as non-adaptive test with only half of the items administered, and at the same time, appropriate items can be selected by the system to measure participant's potential abilities. Therefore, different items will be delivered to different participants in a more time efficient manner.

The data will be analyzed by applying IRT three-parameter logistic (3PL) model. One thousand five hundred and seventy-six participants recruited from Grace Christian Collage in Philippine were administered with Chinese listening and reading tests via CBT in September, 2010. In addition, the effectiveness of applying CAT among the three estimating methods, namely maximum likelihood estimation (MLE), expected a posteriori (EAP), and maximum a posteriori (MAP) will be investigated.

THE COMMON EUROPEAN FRAMEWORK OF REFERENCE, CEFR

CEFR was developed by the Council of Europe (CE) and its members as a framework and guideline for foreign language learning, teaching, and assessment. It was developed as a standard reference and guideline to provide language learning, communication dimension, teaching materials development, and language assessment (Joël Bellassen & Zhang, 2008). CEFR is also a set of language proficiency measurements adapted by different countries to maintain the consistency of mutual authentication between their education systems. The main content of CEFR describes the background of language use, the level of language proficiency, learner acquisition, knowledge, and skills that the language user or learner need to develop (Council of Europe, 2001). CEFR classifies language proficiency and divides proficiency into three categories with a total of six levels (A1, A2, B1, B2, C1, C2). It applies Can-do sentence types and positive presentation types to describe the performance of the various levels of language users and learners' behaviors. Language proficiency as described in CEFR emphasizes the language user and learner's usage of target language in completing certain levels of communication tasks. In order to complete communication tasks, the learner must use their previous experience or competence. Teachers must also understand the ability of the language user and learner in order to support his or her developing ability.

CEFR is an action-oriented approach. It treats language user and learner as part of the community who is able to achieve communication tasks under certain conditions and special circumstances, or some specific behavior aspects (Council of Europe, 2001). Since the 2001 CE recommendation to adopt CEFR, wide spread promotion and application has contributed to the growth of CEFR and has influenced education system in more than 40 countries. Other than CU members countries, countries outside Europe, like Japan, Canada, and New Zealand have referred to CEFR as a framework reference for their foreign language learning, teaching and assessment. Therefore, CEFR is becoming the international language framework reference for language proficiency. Many studies suggest that the most recognized aspect of CEFR is that CEFR has brought positive impact on teaching, curriculum development, and assessment. In the APEC economies research, a survey also showed that CEFR is the best model or reference (Duff, 2008). Therefore, CEFR is a language learning framework that provides clear guidelines for various levels of language learners (Council of Europe, 2001).

Recent Development of Chinese Proficiency Computerized Test

With the popularity of computer devices and the development of information technology, computerized tests have become a current trend in testing. So far, only AP Chinese Language and Culture and TOCFL have developed their own Chinese language computerized assessment systems (College Board, 2011a). AP exams are used for placement purposes to determine college students current language level in the United States. In 2003, the College Board launched an AP Chinese Language and Culture course and exam based on the national standards for foreign language teaching and examination formulated by the American Council on the Teaching of Foreign Languages (ACTFL). The purpose of the exam is to evaluate learners Chinese language communication skills in the real life (College Board, 2011a). TOCFL, on the other hand, is a proficiency test developed for learning CFL by Steering Committee for the Test of Proficiency (SC-TOP) in Taiwan (SC-TOP, 2011). The characteristics of AP Chinese Language and Culture and TOCFL are:

1. There are four kinds of tests in the Chinese language computerized assessment systems, including listening, reading, speaking, and writing tests. The listening and reading tests consist of multiple choice items; the speaking and writing tests comprise free response items.
2. AP Chinese Language exam uses English only for instructions, test content, and computer interface, however, TOCFL uses eight different languages, namely English, Japanese, Korean, French, Spanish, German, Thai, and Vietnamese for instructions, test content, and interface for the Beginner Level, and Chinese for advanced learners.
3. Both test items are presented in simplified or traditional characters. Test takes can choose either Han-Yu-Pin-Yin or Zhu-Yin-Fu-Hao to input their answers.
4. Multiple choice items are scored automatically by the computer system, whereas writing and oral answers are rated manually by Chinese language teachers.

Numerous computerized assessment systems were developed for various language proficiency tests; such as Business Language Testing Service (BULATS) by Cambridge ESOL and Test of English as a Foreign Language (TOEFL), Graduate Record Examination (GRE), and Graduate Management Admission Test (GMAT) by Educational Testing Service (ETS). Among these tests, BULATS and TOEFL were computerized by using CAT (BULATS, 2011; ETS, 2011). Therefore, for Chinese language proficiency tests CAT should be developed and implemented. With the proper use of computer technology, efficiency and accuracy of testing administration could be met by creating more tests with multimedia applications and allowing more flexible testing time.

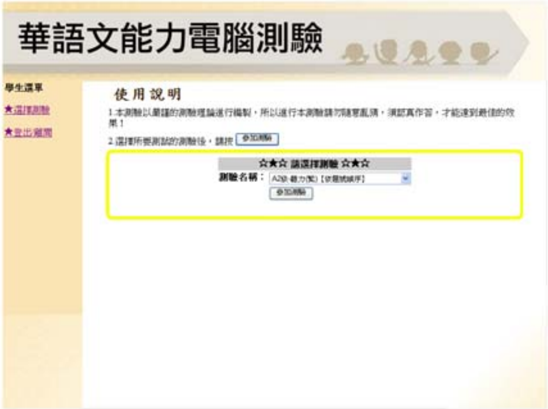
The Development of the CEFR-based Chinese Proficiency Test System

Chinese language proficiency indicators proposed by Tsai (2009) have been adopted in this study and items for the A1 and A2 level of the listening and reading tests have been developed on a web-based test system. In this section, we will introduce the interfaces of the test system, data collection process and the process of developing adaptive testing.

The User Interfaces of Test System

In this system, there are four types of interfaces, test selection, questionnaire, listening test and reading test. These user interfaces are introduced in the following.

Test Selection Interface. Figure 1 indicates the test selection interface. Each examinee has an account number and password which enable them to enter into the system and start the test. Each examinee is required to preselect the section for testing after entering into the system.



華語文能力電腦測驗

學生選單

★選擇測驗

★登出離開

使用說明

1.本測驗以嚴謹的測驗理論進行編製，所以進行本測驗請勿隨意亂填，須認真作答，才能達到最佳的效果！

2.選擇所要測試的測驗後，請按 參加測驗

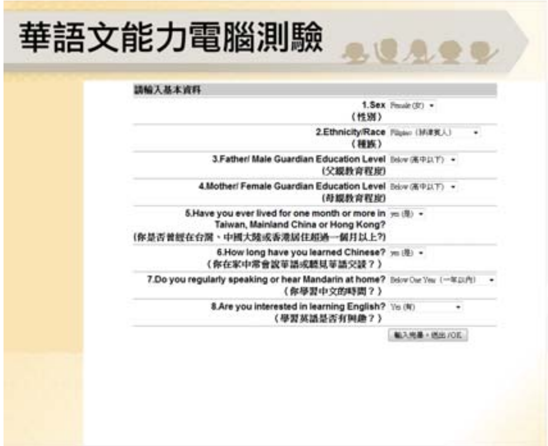
☆★☆ 請選擇測驗 ☆★☆

測驗名稱： A2級-聽力(繁)【依題號順序】

參加測驗

Figure 1. Test Interface

Questionaire Interface. Figure 2 indicates the questionnaire interface. Each examinee have to fill in the basic information questionnaire before the exam starts. The questionnaire is presented in both English and Chinese.



華語文能力電腦測驗

請輸入基本資料

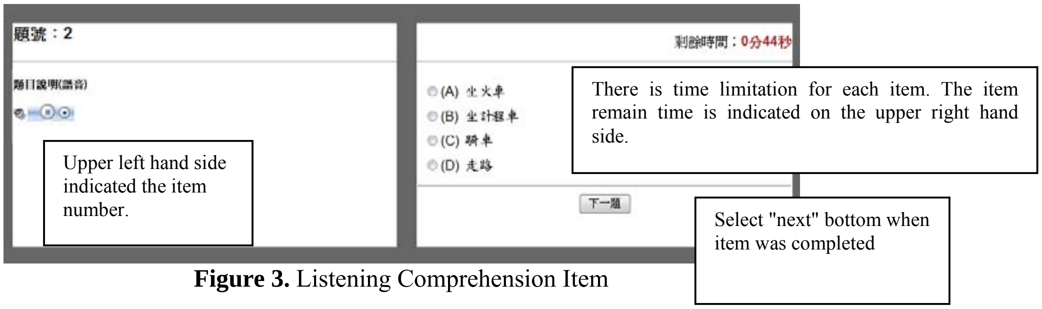
1.Sex （性別）	Female (女)
2.Ethnicity/Race （種族）	Filipino（菲律賓人）
3.Father/ Male Guardian Education Level (父親教育程度)	Below (高中以下)
4.Mother/ Female Guardian Education Level (母親教育程度)	Below (高中以下)
5.Have you ever lived for one month or more in Taiwan, Mainland China or Hong Kong? (你是否曾經在台灣、中國大陸或香港居住超過一個月以上?)	yes (是)
6.How long have you learned Chinese? （你在家中常會說華語或聽見華語交談？）	yes (是)
7.Do you regularly speaking or hear Mandarin at home? （你學習中文的時間？）	Below One Year（一年以內）
8.Are you interested in learning English? （學習英語是否有興趣？）	Yes (有)

輸入完畢，送出 /OK

Figure 2. Questionnaire Interface

The listening test includes listening comprehension and visual-listening comprehension items. In each item, examinees will hear a phrase or a conversation followed by a set of four options to select. Examinees have to click the box that best fits the option on the computer screen. The response time for each item is limited. When the time is up, the system will automatically go to next item.

Listening Comprehension Item: In Figure 3, the examinees will hear, "Walking is too slow! Let us take a taxi to the market. Question: How do they get to the market?" followed with (A), (B), (C), and (D) options The examinees are requested to choose one correct answer. Each item will be read twice and there is a five second break between them.



題號：2

題目說明(語音)

剩餘時間：0分44秒

○(A) 坐火車

○(B) 坐計程車

○(C) 騎車

○(D) 走路

下一題

Upper left hand side indicated the item number.

There is time limitation for each item. The item remain time is indicated on the upper right hand side.

Select "next" bottom when item was completed

Figure 3. Listening Comprehension Item

Visual-Listening Comprehension Item: In Figure 4, the examinees will hear "Please help me buy eggs from the market." and the computer screen will display four options, (A), (B), (C), and (D) on the right hand side.

According to this sentence the examinees have to select an appropriate picture from (A), (B), (C), and (D) which matches the item most. Similarly each item will be read twice.

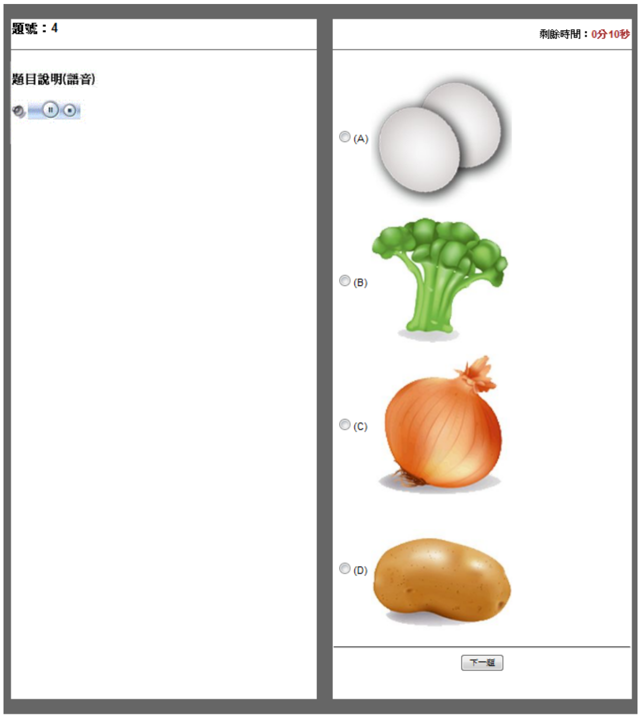
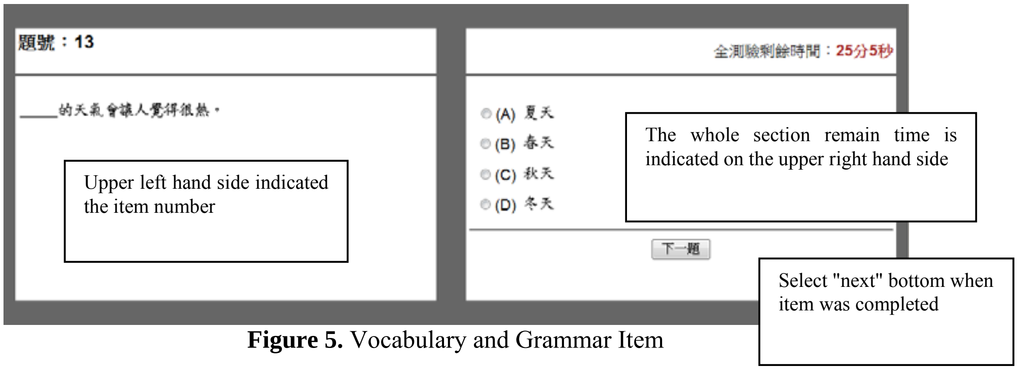


Figure 4. Visual-Listening Comprehension Item

The reading test includes vocabulary and grammar items, and visual comprehension, and reading comprehension items. There is no time limitation on each item in reading test but the whole test has to be completed in 30 minutes.

Vocabulary and Grammar Item: In Figure 5, each item has an incomplete sentence with four options and each option contains a “word” or "vocabulary". Based on this sentence the examinees have to select an appropriate answer which fits the sentence most.



____的天氣會讓人覺得很熱。

○ (A) 夏天
○ (B) 春天
○ (C) 秋天
○ (D) 冬天

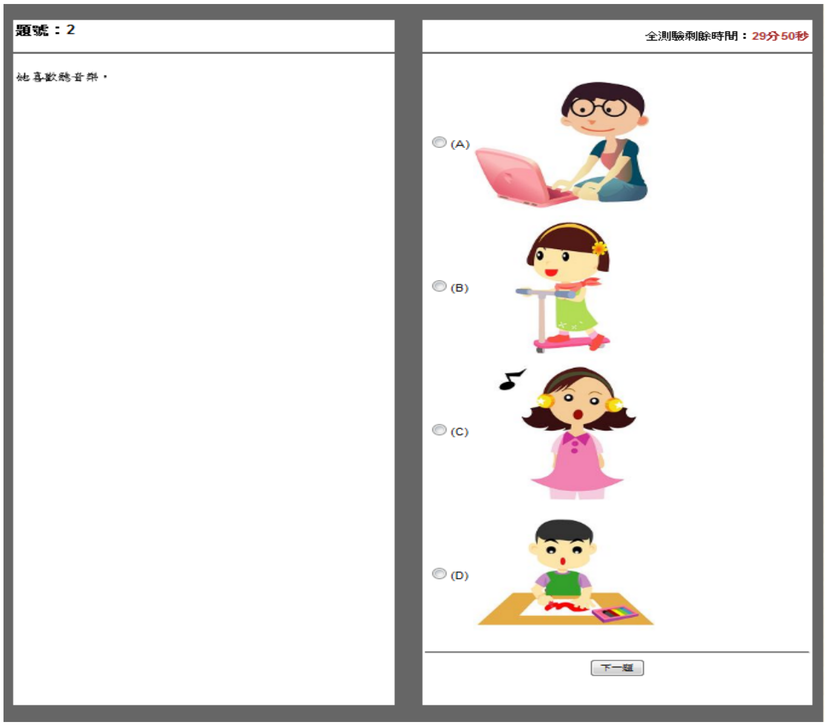
Upper left hand side indicated the item number

The whole section remain time is indicated on the upper right hand side

Select "next" bottom when item was completed

Figure 5. Vocabulary and Grammar Item

Visual Comprehension Item: In Figure 6, the examinees will see a brief sentence, and four options. Based on this sentence the examinees have to select an appropriate picture which matches the item most.



題號：2

她喜歡聽音樂。

全測驗剩餘時間：29分50秒

○ (A)

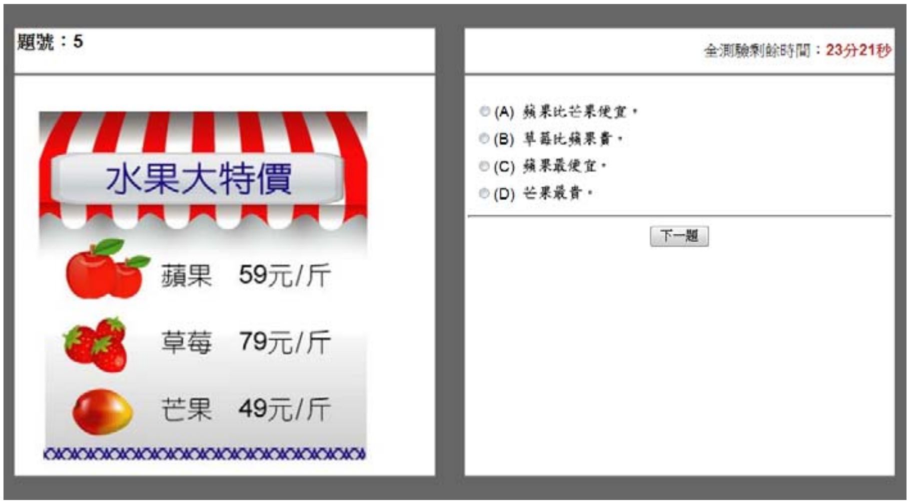
○ (B)

○ (C)

○ (D)

Figure 6. Visual Comprehension Item

Reading Comprehension Item: In Figure 7, the item will include some materials, such as: a picture, flyers, letters, and four options. Based on this information the examinees have to select an appropriate answer from options.



題號：5

水果大特價

蘋果 59元/斤

草莓 79元/斤

芒果 49元/斤

全測驗剩餘時間：23分21秒

○ (A) 蘋果比芒果便宜。

○ (B) 草莓比蘋果貴。

○ (C) 蘋果最便宜。

○ (D) 芒果最貴。

Figure 7. Reading Comprehension Item

Data Collection Process

This study conducted computer-based tests with a listening and reading section for A1 and A2 level exams in the Grace Christian Chinese School in the Philippines. The participants were grades five to ten students. The test level was assigned according to the amount of time the examinee spent learning Chinese. The 5th to 7th grade examinees were assigned to participate in the A1 level CPT. The 5th to 7th grade examinees were assigned to participate in the A2 level. Each examinee must take both listening and reading sections in either the A1 or A2 level. The test time of each test in each level is 30 minutes with a test length of 35. The examinee needs to participate two tests in the same level. Therefore, the total test time for each examinee is 60 minutes.

There were a total of 830 examinees participated in the A1 level and a total of 746 examinees participated in the A2 level. The invalid data were removed including no responses or single responses to the entire set of test items, responses in a guessing manner such as AAABBBCCCDDD for the entire test, response in a very short period time such as responding to each item within 3 seconds, etc.

The Processes of Parameter Estimation and Developing Adaptive Testing

Many CPTs are still using the raw score to categorize proficiency level. For example, HSK transfers the exam's raw score as the test result (HSK, 2011). The score report in the SAT Chinese test includes the raw score, composite total score, and percentile range (NCACLS, 2010). The score calculation in TOCFL is based on one point for each correct item and there is no penalty for wrong answers (SC-TOP, 2011). Wright (1999) shows that if using raw score to represent proficiency, the test is not able to become a meaningful measurement because of the lack of virtue in basic requirements such as unidimensionality. Lord (1980) pointed out that IRT had improved those shortcomings in CTT such as the assumption of using a single standard of error measurement, sample dependent parameter estimation, parallel tests assumptions. Therefore, this study applies a 3PL IRT model for item and ability parameter estimation. The resulting 3PL model (Baker, 1992; Baker & Kim, 2004; Zimowski, Muraki, Mislevy, & Bock, 2003) is

$$P(x_j = 1 \mid \theta_k, a_j, b_j, c_j) = c_j + \frac{(1-c_j)}{1+\exp^{-D^{*}a_j(\theta_k - b_j)}} \equiv P_{j1}(\theta_k) \qquad (1)$$

where $P_{j1}(\theta_k)$ is the probability that an examinee with ability θ_k answers item j correctly; a_j is the item discrimination for item j; b_j is the item difficulty for item j and b_j represents the point on the ability scale at which a candidate has a 50% probability of answering item j correctly; c_j is the item guessing for item j; D is a scaling factor and is applied the default value, 1.7.

The marginal maximum likelihood (MMLE) formulation with an expectation-maximization (EM) algorithm is applied to calibrate the item and ability parameters (Zimowski, Muraki, Mislevy, & Bock, 2003). An item bank was established after obtaining the item parameters. One of the goals of this study is to develop a computerized adaptive test for the CPT.

Figure 8 shows the structure of an adaptive test as a flowchart in this study. The three major steps (starting, continuing, and stopping) were followed the flowchart. The steps were (Wainer, 2000):

Starting: The general principle of selecting the next item based on previous response is not helpful, of course, when there are no previous responses. Although an examinee's proficiency cannot be estimated from responses to previously administered items when testing begins, the mean of the population of examinees is a reasonable initial guess. After a few response, examinees lead themselves to items that are more informative near their own particular.

Continuing: The two strategies currently most widely used for selecting an examinee's next item, given a provisional estimate of ability based on preceding responses, are methods providing "maximum information" and "maximum expected precision". In this study item selection strategy is based on the maximum information method. The item selection procedure is the process of selecting an item from the item pool to be administered to the examinee, and that information will be provided as a guideline in the CAT system to indicate which items should or should not be chosen during a test.

Stopping: After each item is administered and scored, an interim estimate of the examinee's ability is calculated and used by the item selection procedure to select the next item. Three commonly used ability estimation procedure are MAP, MLE, and EAP (Lord, 1980). An adaptive test can be terminated when a target measurement precision has been attained, when a preselected number of items has been given, or when a predetermined amount of time has elapsed. Any of these rules may be used in its pure form, or a mixture of them can be used.

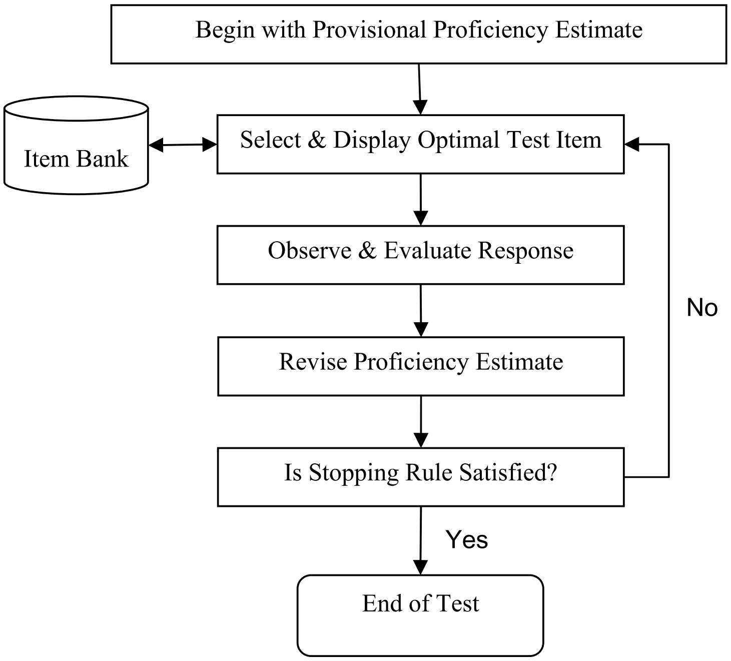

Begin with Provisional Proficiency Estimate
Item Bank
Select & Display Optimal Test Item
Observe & Evaluate Response
No
Revise Proficiency Estimate
Is Stopping Rule Satisfied?
Yes
End of Test

Figure 8. The Procedure of Computerized Adaptive Testing

RESULTS

The Reliability and Item Parameters of Chinese Proficiency Test

Table 1 shows the reliability (Cronbach's α) of the Chinese proficiency test. The reliability in each section of the test ranged from 0.842 to 0.899 reflect a reasonable degree of reliability. In general, an alpha value of around 0.8 is an acceptable value for Chinese proficiency test. (Shen, 2005), this means that these Chinese proficiency tests are reliable.

Table 1. **The Test Reliability**

Section	Test Length	Effective Sample Size	Cronbach's α
A1 listening	35	798	0.842
A1 reading	35	797	0.897
A2 listening	35	712	0.899
A2 reading	35	706	0.869

The averages of item parameters for each section presented in Table 2 show that the average of item discrimination in the listening and reading sections were higher than 1.2. This indicated a very high degree of item discrimination had developed in each section. In addition, according to the IRT model, the average correct rates are 68.44%, 69.51%, 69.38%, and 58.26% for A1 listening, A1 reading, A2 listing, and A2 reading section respectively.

Table 2. **Averages of Item Parameters in Each Section**

Section	a	b	c	P(Θ)
A1 listening	1.2223	-0.4742	0.2075	0.6844
A1 reading	1.2425	-0.4473	0.2048	0.6951
A2 listening	1.3145	-0.4637	0.1998	0.6938
A2 reading	1.2125	-0.0214	0.2085	0.5826

The Chinese Proficiencies of Total, Male and Female Groups

Table 3 shows the sample sizes of the total group and the gender subgroups for each of the 4 forms of the CPT. Over the various test forms, the male group comprised 47% to 48% of the total group, and the female group comprised 52% to 53% of the total group. The average number-correct scores and standard deviations for groups taking different forms of the CPT are summarized in Table 4. It shows that the female group had higher mean scores than the male group. The average raw scores across various test forms were similar to one another, both for the total group and for each of the gender subgroups. This provided evidence of random assignment of test forms to candidates (i.e., the groups taking different forms were fairly equivalent). Overall, Table 4 shows that the test forms were designed to be fairly similar to one another.

Table 3. **Sample Sizes of Total and Gender Subgroups on CPT**

Section	Total Group (n)	Male Group		Female Group	
		n_m	$\frac{n_m}{n}$	n_f	$\frac{n_f}{n}$
A1 listening	798	381	0.48	417	0.52
A1 reading	797	379	0.48	418	0.52
A2 listening	712	337	0.47	375	0.53
A2 reading	706	333	0.47	373	0.53

Note. n The sample sizes of the total group.
n_m The sample sizes of the male group.
n_f The sample sizes of the female group.

Table 4. **Average Raw Scores of Total Group and Gender Subgroups on CPT**

Section	N	Total Group		Male Group		Female Group	
		M	SD	M	SD	M	SD
A1 listening	798	23.93	6.31	22.50	6.61	25.25	5.72
A1 reading	797	24.27	7.07	22.21	7.46	26.13[a]	6.12
A2 listening	712	24.23	6.96	22.62	7.52	25.67	6.07
A2 reading	706	20.47	6.83	19.19	7.20	21.62[b]	6.28

Note. a. The maximum of means.
b. The minimum of means.

The Effectiveness of CAT System for CPT

In this study, a complete computerized test without adaptive process was applied to collect participants' responses. And these responses were used to estimate the items parameters and evaluate the performances of different ability estimation methods in CAT process. The evaluation method is applied the collected data into CAT process mentioned in Figure 8 to simulate CAT process. At each iteration, CAT assumes one item is draw from item bank and administered to the participant. We can obtain the response of this item in the collected data.

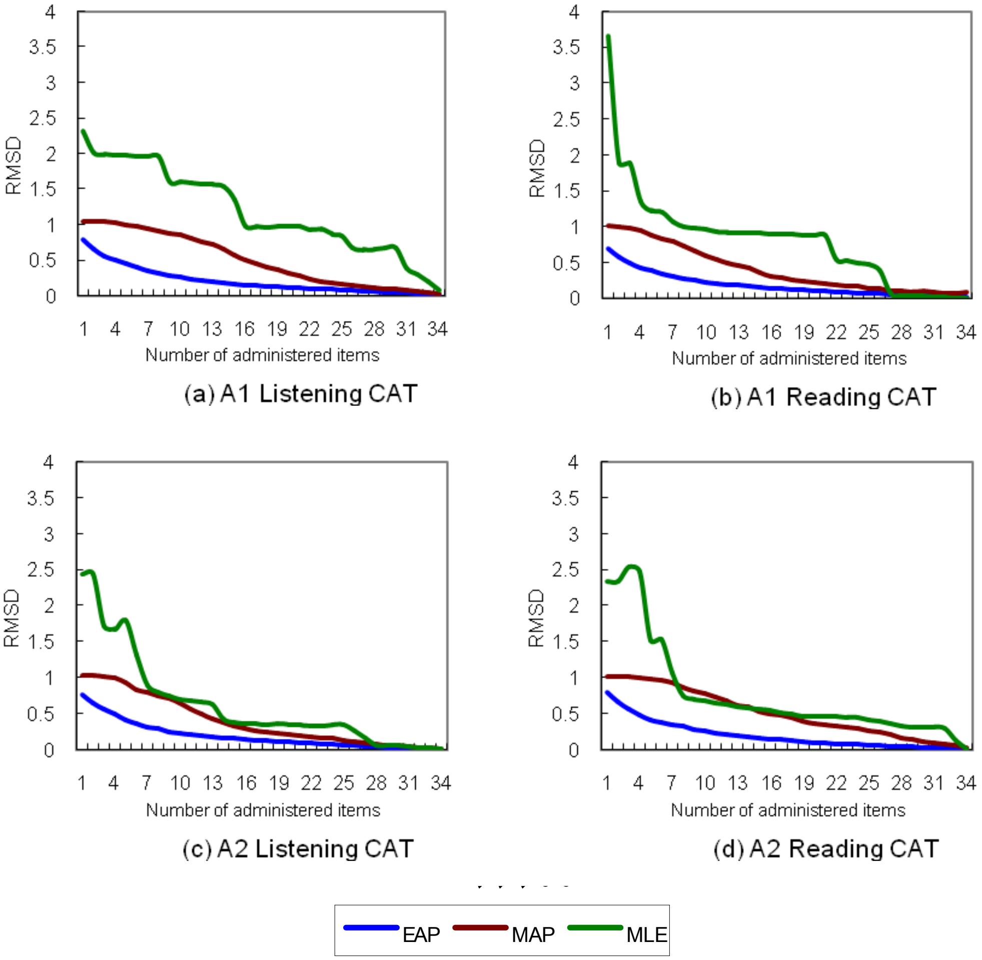
For evaluating the performances of CAT algorithms based on different ability estimation methods, MLE, MAP and EAP, the root mean squared difference (RMSD) between the estimated abilities by CAT and by complete test was applied. The definition of RMSD is stated in following

$$RMSD\ (\hat{\theta}_i^{(k)}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_i^{(k)} - \tilde{\theta}_i)^2}$$

(2)

where $\tilde{\theta}_i$ represents the i^{th} participant's ability estimated by using all administrated items. $\hat{\theta}_i^{(k)}$ represent the i^{th} participant's temporarily ability estimate after k items had been responded (in Kth iteration); N represents the total number of participants.

In Figure 9, the vertical axis indicates the RMSDs of EAP, MAP, and MLE and the horizontal axis represents the number of administered items. Figure 9 shows that there is a significant difference in RMSD decline as the accumulation of items examinees participated in increased. Referring to the estimated result from Figure 9a; it indicated that, using MLE, the RMSDs are greater than 1 when exam items completed number less than 15 and the RMSDs are less than 0.4 when the exam items completed reached 31. In addition, when using MAP, the RMSDs are greater than 1 when exam items completed number less than 5 and the RMSDs are less than 0.4 when the exam items completed reached 19. However, when using EAP, the RMSDs are always less than 1 and the RMSDs are less than 0.4 when the exam items completed reached 6. The other three sections also showed the same result regarding these three estimation methods. This result indicated that under above three estimation methods, the EAP estimation method resulted in an overall lower RMSD compared with MLE and MAP. This result is similar to the study conducted by Chen (2006), Wang and Vispoel (1998). Therefore, the EAP parameter estimation method was adopted in the proposed CAT system.



(a) A1 Listening CAT

(b) A1 Reading CAT

(c) A2 Listening CAT

(d) A2 Reading CAT

Figure 9. The Performances of EAP, MAP, and MLE in CAT

DISCUSSION

This study based on CEFR and Tsai (2009) developed A1 and A2 level items for a CPT in both listening and reading sections. The computerized CPT was performed onsite at the Grace Christian Chinese School in the Philippines. The examinees were 5th to 10th grade CFL learners. The development of the CPT in this study refers to the PISA 2006 test development process (OECD, 2009). The data analysis showed that the computerized CPT possess good reliability and validity. The examinee's item correct response rate in different tests is close to 70% except for the A2 level reading section which is closer to 60%. The results also indicated that females performed better than males.

The CAT system developed in this study included a testing interface and a management interface. For the testing interface, examinees participated in testing according to their proficiency level after login to the interface. The result will also be presented to the examinee as soon as the items are competed. The management interface contains the function of item bank editing. This function also includes test assignment, item bank creation or modification, and item editing in the item bank. In addition, there are different features in the CAT system that are available to the user in accordance to his or her requirement. For example, the user can select different testing formats and different parameter estimation methods. In response to international demand, the CAT system for CPT developed in this study used computer facilities to analyze and calibrate the test and score. This will shorten the data collection time. When performing the CAT simulation through different parameter estimation methods, this study discovered that the RMSD is best performed under the EAP estimation method. Therefore, this study recommends EAP as the prefered parameter estimation method.

During research, valuable experience was acquired during the system implementation process and the actual conduct of the test. This valuable experience can be used as directions in future research and subsequent recommendations are as follows:

1. This CAT system was developed for multiple-choice items. However, in order to fully utilize computers in the test, this CAT system can be amended to fit more diverse and comprehensive items and to make the exam closer to real scenarios.
2. The extension of this study is to develop the B level or even the C level of the CPT and focus on new item format development in the near future, not only to enrich and enliven the content of the CPT but also to be able to implement proficiency test according to the examinee's ability in productive activities and strategies, receptive activities and strategies, interactive activities and strategies, and mediating activities and strategies.
3. Considering the examinee's proficiency and acceptability, future studies can focus more on conducting the test or grading online with a CAT system for the writing and speaking section to make it more common and easy to carry out the assessment.
4. The CAT system was developed based on traditional Chinese. In consideration of the majority of the users and learners in different countries, a simple Chinese version can also be implemented rendering the test limitless cross the world.
5. This study can also focus on adding new functions to the CAT system such as the an initial item setup method, item selection strategy, and exposure rate control, in the near future.

ACKNOWLEDGEMENTS

This work was supported in part by the National Science Council, Taiwan, under Contracts NSC 99-2410-H-142 -008 -MY3 and NSC 101-2631-S-003-001.

REFERENCES

Akdemir, O., & Oguz, A. (2008). Computer-based testing: An alternative for the assessment of Turkish undergraduate students. *Computers & Education, 51*(3), 1198-1204.

Baker, F. B. (1992). *Item Response Theory: Parameter Estimation Techniques.* New York: Marcel Dekker.

Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York: Marcel Dekker.

Birnbaum, A. (1968). *Some latent trait models and their use in inferring an examinee's ability*. In F. M. Lord & M. R. Novick (Eds.), Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

BULATS (2011). *Business Language Testing Service.* Retrieved June 17, 2011, from http://www.bulats.org/

C. Test (2011). *Test of Practical Chinese*. Retrieved May 27, 2011, from http://www.c-test.org.cn/

Chen, P. H. (2006). The Influences of the Ability Estimation Methods on the Measurement Accuracy in Multidimensional Computerized Adaptive Testing. *Bulletin of Educational Psychology, 38(2),* 195-211.

College Board (2011a). *Chinese with Listening*. Retrieved May 20, 2011, from http://www.collegeboard.com/student/testing/sat/lc_two/chinese/chinese.html?chinese

College Board (2011b). *Chinese language and culture*. Retrieved May 7, 2011, from http://www.collegeboard.com/student/testing/ap/sub_chineselang.html

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge, UK: Cambridge University Press.

Duff, P. A. (2008). *Foreign Language Policies, Research, and Educational Possibilities.* APEC Education Symposium, Xi'an, China, 14-17 January 2008. Available at: http:// www.sei2003.com/APEC/LearningStandards/Duff_APEC_2008.doc.

ETS (2011). *The TOEFL test.* Retrieved June 17, 2011, from http://www.ets.org/toefl

HSK (2011). *Hanyu Shuiping Kaoshi*. Retrieved May 21, 2011, from http://www.hsk.org.cn/index.aspx

Jodoin, M. G. (2003). Measurement efficiency of innovative item formats in computer-based testing. *Journal of Educational Measurement, 40* (1), 1-15.

Joël Bellassen & Zhang, L. (2008). The enlightenment and the impetus of the new approach of the Common European Framework of Reference for Language on the Chinese language teaching. *Chinese Teaching in the World*, (3), 58-73

Kecker, G., & Eckes, T. (2007). *Linking the TestDaF to the CEFR: The case of writing proficiency.* Fourth Annual EALTA Conference, June 15-17, 2007.

Lord, F. M. (1980). *Application of item response theory to practical testing problems*. hillsdale, NJ : lawrence erlbaum associates.

Mills, J. D. (2002). Using computer simulation methods to teach statistics: A review of the literature. *Journal of Statistics Education, 10*(1), http://www.amstat.org/publications/jse/v10n1/mills.html.

NCACLS (2010). *National Preparation Test for SAT Subject Test in Chinese with Listening*. Retrieved May 7, 2010, from http://www.scccs.net/events/event34/SATII/2010SATII.pdf

NCACLS (2010). *National Preparation Test for SAT Subject Test in Chinese with Listening*. Retrieved July 13, 2010, from http://www.scccs.net/events/event34/ SATII/2010SATII.pdf

OECD (2009). *PISA 2006 Technical Report*. OCED, Paris.

SC-TOP (2011). *Steering Committee for the Test of Proficiency-Huayu*. Retrieved May 17, 2011, from http://www.sc-top.org.tw/

Shen, H. H. (2005). An investigation of Chinese-character learning strategies among non-native speakers of Chinese. *System, 33*(1), 49-68.

Shieh, J. C. (2011). The Unified Phonetic Transcription for Teaching and Learning Chinese Languages. *Turkish Online Journal of Educational Technology, 10(4)*, 355-369.

Tannenbaum, R. J., & Wylie, E. C. (2005). *Mapping English Language proficiency test scores onto the common European framework.* (ETS Research Rep. No. RR-05-18; TOEFL Research Rep. No. RR–80). Princeton, NJ: ETS.

Tsai, Y. H. (2009). CFL Teaching Materials Construction. Taipei: Zhong Zheng.

U.S. Department of State (2006). *National Security Language Initiative*. Retrieved May 21, 2011, from http://merln.ndu.edu/archivepdf/nss/state/58733.pdf

Wainer, H. (2000). *Computerized adaptive testing: A primer* (Second Edition). Hillsdale, NJ: Lawrence Erlbaum Associates.

Wang, T. & Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement, 35*, 109-135.

Wise, S.L., & Plake, B.S. (1990). Computer-based testing in higher education. *Measurement and Evaluation in Counseling and Development, 23*(1), 3–10.

Wong, L. H., Gao, P., Chai, C.S., & Chin. C. K. (2011). Where Research, Practice and the Authority Meet: A Collaborative Inquiry for Development of Technology-Enhanced Chinese Language Curricula. *The Turkish Online Journal of Educational Technology, 10 (1)*, 232-243.

Wright, B. D. (1999). *Fundamental measurement for psychology.* The new rules of measurement. S. E. Embretson and S. L. Hershberger. Mahwah NJ, Lawrence Erlbaum Associates.

Zhao, X. L., Wang, M. J., Wu, J., & He, K. K. (2011). ICT and An Exploratory Pedagogy for Classroom-Based Chinese Language Learning. *The Turkish Online Journal of Educational Technology, 10 (3)*, 141-151.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG*. Scientific Software International.