The Development and Evaluation of a Computerized Adaptive Testing System for Chinese Proficiency - Base on CEFR

Hsuan-Po Wang\textsuperscript{a}, Bor-Chen Kuo\textsuperscript{a}, Rih-Chang Chao\textsuperscript{a}, Ya-Hsun Tsai\textsuperscript{b}

\textsuperscript{a}Graduate Institute of Educational Measurement and Statistic, National Taichung University of Education, 140 Min-Sheng Rd., Taichung, Taiwan 40306, R.O.C.

\textsuperscript{b}College of International Studies and Education for Overseas, National Taiwan Normal University, 162 Hoping East Road, Sec. 1 Taipei, Taiwan 106, R.O.C.

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.

© 2012 Published by Elsevier Ltd. Selection and/or peer-review under responsibility of The Association Science Education and Technology Open access under CC BY-NC-ND license.

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

1. 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. 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.

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 College in Philippine were administered with Chinese listening 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.

2. 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). 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). 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).

3. 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 test 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.

3.1. The User Interfaces of Test System
In this system, there are three types of interfaces, test selection, questionnaire, and listening test. These user interfaces are introduced in the following.

Fig. 1.(a) 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. Fig. 1.(b) indicates the questionnaire interface. Each examinee has to fill in the basic information questionnaire before the exam starts. The questionnaire is presented in both English and Chinese.

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 Fig. 2, 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.

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.
Visual-Listening Comprehension Item: In Fig. 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.

Fig. 2. Listening Comprehension Item

Fig. 3. Visual-Listening Comprehension Item

4. Data Collection Process

This study conducted computer-based tests with listening 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. The test time of each test in each level is 30 minutes with a test length of 35.

5. The Processes of Parameter Estimation and Developing Adaptive Testing

This study applies a 3PLIRT model for item and ability parameter estimation. The resulting 3PL model (Baker, 1992; Baker & Kim, 2004; Zimowski, Muraki, Mislevy, & Bock, 2003) is

\[
P(x_j = 1 | \theta_i, a_j, b_j, c_j) = c_j + \frac{(1 - c_j)}{1 + \exp(-D(a_j \theta_i - b_j))} = P_j(\theta_i)
\]

where \(P_j(\theta_i)\) is the probability that an examinee with ability \(\theta_i\) 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.

Fig. 4 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):

![Flowchart of Computerized Adaptive Testing](image)

6. Results

6.1. The Reliability and Item Parameters of Chinese Proficiency Test

Table 1 shows the reliability (Cronbach's $\alpha$) 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>
<thead>
<tr>
<th>Level</th>
<th>Test Length</th>
<th>Effective Sample Size</th>
<th>Cronbach’s $\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>35</td>
<td>798</td>
<td>0.842</td>
</tr>
<tr>
<td>A2</td>
<td>35</td>
<td>712</td>
<td>0.899</td>
</tr>
</tbody>
</table>

The averages of item parameters for each section presented in Table 2 show that the average of item discrimination in the listening section 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% and 69.38% for A1 and A2 listening section respectively.
Table 2. Averages of Item Parameters in Each Section

<table>
<thead>
<tr>
<th>Level</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>P(θ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>1.2223</td>
<td>-0.4742</td>
<td>0.2075</td>
<td>0.6844</td>
</tr>
<tr>
<td>A2</td>
<td>1.3145</td>
<td>-0.4637</td>
<td>0.1998</td>
<td>0.6938</td>
</tr>
</tbody>
</table>

6.2. 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 2 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

<table>
<thead>
<tr>
<th>Level</th>
<th>Total Group (n)</th>
<th>Male Group</th>
<th></th>
<th>Female Group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total Group (n)</td>
<td>Male Group</td>
<td>Female Group</td>
<td></td>
</tr>
<tr>
<td></td>
<td>n_m</td>
<td>n_f</td>
<td>n_m/n</td>
<td>n_f/n</td>
</tr>
<tr>
<td>A1</td>
<td>798</td>
<td>381</td>
<td>0.48</td>
<td>417</td>
</tr>
<tr>
<td>A2</td>
<td>712</td>
<td>337</td>
<td>0.47</td>
<td>375</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Level</th>
<th>N</th>
<th>Total Group</th>
<th>Male Group</th>
<th>Female Group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>A1</td>
<td>798</td>
<td>23.93</td>
<td>6.31</td>
<td>22.50^a</td>
</tr>
<tr>
<td>A2</td>
<td>712</td>
<td>24.23</td>
<td>6.96</td>
<td>22.62</td>
</tr>
</tbody>
</table>

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

6.3. 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 Fig.4 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

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

where \(\hat{\theta}_i\) 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 \(k^{th}\) iteration); \(N\) represents the total number of participants.

In Fig.5, the vertical axis indicates the RMSDs of EAP, MAP, and MLE and the horizontal axis represents the number of administered items. Fig.5 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 section 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.

Fig.5. The Performances of EAP, MAP, and MLE in CAT
7. Discussion

This study, based on CEFR and Tsai (2009), developed A1 and A2 level items for a CPT in the listening section. 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 possesses good reliability and validity. The examinee’s item correct response rate in different tests is close to 70%. 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 preferred 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 across 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.

References


