International Journal of Innovative
Computing, Information and Control
Volume **11**, Number **1**, February **2015**

ICIC International ©2015 ISSN 1349-4198
pp. 57–76

DEVELOPMENT OF CHINESE COMPUTERIZED ADAPTIVE TEST SYSTEM BASED ON HIGHER-ORDER ITEM RESPONSE THEORY

Rih-Chang Chao[1], Bor-Chen Kuo[1] and Ya-Hsun Tsai[2]

[1]Graduate Institute of Educational Measurement and Statistics Department
National Taichung University
No. 140, Min-Shen Road, Taichung 40306, Taiwan
rchang.chao@msa.hinet.net; kbc@mail.ntcu.edu.tw

[2]Applied Chinese Language and Culture Department
National Taiwan Normal University
No. 162, Section 1, Heping East Road, Taipei 10610, Taiwan
yahsun@ntnu.edu.tw

Received February 2014; revised June 2014

Abstract. *The purpose of this empirical study was to develop a higher-order computerized adaptive test (HCAT) system based on higher-order Item Response Theory (HO-IRT) for a Chinese proficiency test (CPT). This study contains four steps, namely item categorization, ability estimation, item selection and system development. The development of the HCAT system succeeds in making four major contributions; the first is the original CPTs system developed for computerized adaptive test, while the second is the application of one factor within-item HO-IRT structure which enables the item selection procedure adopted next item simultaneously based on the individual test-taker's three domain abilities (linguistic, sociolinguistic and pragmatic), thirdly is the application of the maximum a posteriori (MAP) estimator enables test-takers averagely responded 10 out of 43 items to complete the test, and last, but not least, the HCAT system is able to directly reduce the administration costs associated with the test compared with either the current traditional paper-and-pencil or computer-based CPTs. These results make CPT a multi-functional and practical test, as the domain abilities can be used in a formative assessment for diagnostic purposes, and the overall ability can be provided on the overall level of performance for a summative assessment.*

Keywords: Item response theory (IRT), Higher-order IRT (HO-IRT), Computerized adaptive test (CAT), Higher-order computerized adaptive test (HCAT), Chinese proficiency test (CPT)

1. **Introduction.** Language skills have traditionally been distinguished in terms of channel (audio, visual) and mode (productive, receptive), and recognized in the form of four skills, namely, listening, reading, speaking, and writing [1,2]. Currently the Chinese Proficiency Tests (CPTs) measure communicative language ability (CLA) and calibrate test-takers' ability via the language skill approach. For example, the CPTs performed in the USA (The Advanced Placement Chinese Language and Culture Exam), China (Hanyu Shuiping Kaoshi), and Taiwan (Test of Chinese as a Foreign Language), all focus on measuring of test-takers' abilities in these four skills.

Bachman and Palmer [3] argue that the notion of a language skill approach is unsuitable for using in language testing because it does not reveal the differences between language use activities that are considered to be within the same "skill". In addition, the most important revelation derived from a further analysis of related research studies is that CLA consists of a general higher-order factor [4-7].

A number of researchers have endorsed the fact that CLA is multi-componential [8-10]. Sawaki, Sticker and Oranje [4] point out that the multidimensional CLA comes in different forms that vary in terms of the exact factor structure identified in the method of confirmatory factor analysis (CFA). However, according to Stone and Yeh [11], the application of CFA on dichotomous data can be problematic. Firstly, the item distribution differences in the test administration and low reliable responding data from test-takers lead to spurious results (e.g., Ackerman, Gierl and Walker, 2003; Green, 1983; Swygert, McLeod and Thissen, 2001) [12-14], in which the factor loading will be underestimated and the number of dimensions overestimated (e.g., Bock, Gibbons and Muraki, 1988) [15]. Finally, the estimated error will be enlarged because the guessing parameter has increased the probability of responding to multiple choice items [16]. Although researchers are not able to perspicuously elaborate on the supposed components of CLA or what should be interactively involved in the relationship among those components [17-22], nevertheless, a consistent consensus is reached that CLA is multidimensional [4,23], and that the components of CLA should be measured by means of both separated and integrated tasks [1]. For example, the TOEFL iBT has measured language skill components through both separated and integrated task items [24]. In comparison, the framework of the current CPTs does not include a higher-order ability. Furthermore, none of them calibrate test-takers' linguistic, sociolinguistic, and/or pragmatic ability through both separated and integrated task items.

Lord [25] initially studied flexilevel testing, which led to a boom in the development of the computerized adaptive test (CAT) system [26]. The CAT system is also called a tailored test, since the items selected for test-takers to respond to are based on their individual provisional ability. This enables an accurate estimate to be made of the test-taker's ability by adopting fewer items than either the traditional paper-and-pencil or computer-based test [27,28]. However, the advantages of CAT technology had been recognized and utilized for years, but none of CPTs adopt a CAT system. Therefore, the goal of this study is to develop a higher-order CAT system through both separated and integrated task items, which enables the simultaneous measurement of the test-takers' reading ability and their linguistic, sociolinguistic, and pragmatic abilities in Chinese communicative language proficiency test.

2. **Higher-Order Computerized Adaptive Test System.**

The test framework of Item Response Theory (IRT) is a mathematical structure based on individual test-takers' responses to items and it is adopted according to the characteristics of each item (e.g., applied difficulty, discrimination and guessing parameters) and the test-taker's ability or latent traits. There are two structures to differentiate test-takers' ability; one of which is the unidimensional IRT (UIRT), while the other is the multidimensional IRT (MIRT). One of the major assumptions of the UIRT is that the ability estimated is one-dimensional, which means that all the items administered in the test are accounted for by unidimensional ability. When the multiple abilities of the test-taker are correlated and to be measured, the results of the parameter estimation will not be accurate when applying a UIRT because it is incapable of simultaneously estimating items accounted for by abilities of two or more dimensions. Consequently, as indicated by Ackerman, Gierl and Walker, the ability will be overestimated when using a higher discrimination parameter, or the ability will be underestimated (or even neglected) when using a lower discrimination parameter [12]. The MIRT structure was developed to handle and manipulate these shortcomings [29-33], and its main function is to estimate the abilities of test-takers on a multidimensional basis.

The CAT system can be distinguished by a unidimensional CAT (UCAT) system and multidimensional CAT (MCAT) system based on the abilities of the test-taker being measured. The difference between these systems is that the latter is able to simultaneously measure test-takers' multiple abilities, while the former needs to measure multiple abilities separately. When the test-takers' multiple abilities are correlated and each administrated item is accounted for by one dimensional ability, this is the result of the MCAT system, while its test framework was based on a between-item structure, are more accurate than those of the UCAT system. This is because it is not only able to estimate all abilities simultaneously, but at the same time, the correlations of multidimensional abilities are utilized and input to its estimation procedure [34-36].

Segall [35] discovered that the MCAT system administered about one-third fewer items than the UCAT system. However, Segall and Moreno [37] adopted empirical data and unidimensional item parameters, which were estimated from a computer-based test, to conduct a simulated study of the MCAT system. This simulation result is considered problematic because if the administered item is accounted for by two or more domain abilities and the test-taker lacks one or some of those domain abilities, the guessing parameter will be considered to have been the reason the test-taker responded correctly to this item. The accuracy of the calibrated abilities is then become uncertain. Furthermore, the MCAT system is incapable of estimating test-takers' overall ability. Therefore, there is a tendency to apply a higher-order CAT system to assess test-takers with an overall ability [32,33,38,39].

The higher-order CAT system, which adopts a linear mathematical structure, one-factor higher-order IRT (HO-IRT) structure, is not only able to simultaneously estimate test-takers multidimensional domain abilities, but is also capable of assessing their higher-order ability by utilizing correlations [40-42]. The one-factor HO-IRT structure comprises two distinctive test frameworks of between-item and within-item [43]. The former assumes that each item is accounted for by unidimensional ability. For example, Huang et al. [44] had successfully assessed test-takers' domain abilities and overall ability by conducted a simulated study of higher-order CAT system based on a one factor between-item HO-IRT structure. The one factor within-item HO-IRT structure assumes that some items are simultaneously accounted for by two or more abilities. It is normal to find out that in some situations, the test-takers with a low level reading ability can perform similar to those with an intermediate level of reading ability. For example, the more knowledge of the Chinese culture and history test-takers acquired, the better their comprehension of the meaning behind vocabulary and the higher score they can achieve in the CPT. Since the test framework of this study needs to comprise linguistic, sociolinguistic and pragmatic ability and the items implemented in this study had to include both separated and integrated task items. Therefore, the framework of one factor within-item HO-IRT structure was adopted in this empirical study for the development of the CAT system (HCAT system).

3. **Procedures.**

The procedures of this study contain four steps; the first of which is to conduct the item categorization. The second step is the mathematical formula derivation undertook for the domain and overall abilities estimation, while the third step is the mathematical formula derivation undertook for item selection procedure. The fourth step is to develop the HCAT system.

3.1. **Item categorization.**

The items of this study were implemented by adopting PIS A's (The Programme for International Student Assessment) standard procedures, which

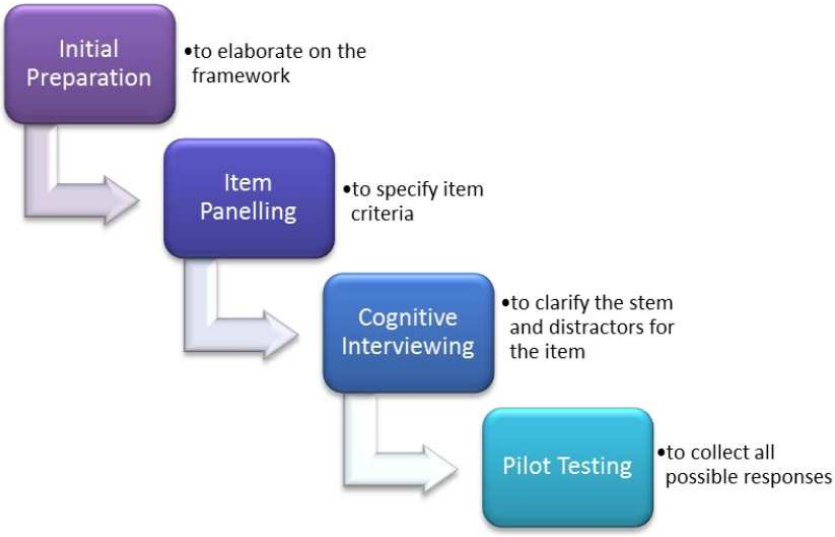

FIGURE 1. Implementation procedures

are included in four steps shown in Figure 1 [45], namely, initial preparation, item panelling, cognitive interviewing, and pilot testing. These items were implemented by professors of the Department of Applied Chinese Language and Culture at the National Taiwan Normal University. All the items were thoroughly reviewed based on the item writing expertise and judges' consensus to ensure that they were appropriate before they were integrated into the test items.

The Common European Framework of Reference for languages (CEFR) was developed based on what learners need to know and what they should be able to achieve in the target language [46]. It is components of CLA are derived from the opinions of experts with first-hand experience of teaching the local curriculum and assessing students' performance [23]. CEFR provides some useful descriptive scales with a broad compendium of information about the consensus views of language learning, teaching, and assessment. The proficiency descriptors in the CEFR specify the level of CLA comprised linguistic, sociolinguistic, and pragmatic ability within a language skill. Linguistic ability consists of lexical, grammatical, semantic, phonological, and syntactical ability, while sociolinguistic ability relates to the knowledge and skills required to deal with the social dimension of language use. Pragmatic ability involves the functional use of linguistic resources, including discourse and functional ability.

The items were categorized to correspond to linguistic, sociolinguistic, or/and pragmatic ability based on the descriptor scales of CEFR. These items can be categorized into three different types according to the numbers of domain abilities for which they are accounting, as shown below.

A. Category I: Items that are accounted for by single domain ability. There are seven items, each of which is accounted for by either linguistic, sociolinguistic or pragmatic ability. All the items in category I are separated task items.

B. Category II: Items that are accounted for by two domain abilities. There are five items, each of which is accounted for by a combination of two domain abilities (linguistic and sociolinguistic, linguistic and pragmatic, or sociolinguistic and pragmatic). All the items in category II are integrated task items.

C. Category III: Items that are accounted for by three domain abilities. There are seven items, each of which is accounted for by a combination of linguistic, sociolinguistic and pragmatic ability. All the items in category III are integrated task items.

Item 26 below is an example of how the items are categorized as a separated task item.

Item 26

小明：小強的溝通能力如何？

Xiao-Ming: What is the CLA of Xiao-Quang?

小張：跟他聊天時非常愉快，他總是「侃侃而談」（kǎnkǎnértán）。

Xiao-Zhang: It is pleasant to have a conversation with him; he is always kǎnkǎnértán [eloquent]; the conversation is not limited to any one topic.

「侃侃而談」表示：

What is the meaning behind the words, kǎnkǎnértán [eloquent]?

(A) 胡言亂語 húyán-luànyǔ [nonsense]

(B) 對答如流 duìdárúliú [fluent]

(C) 陳腔濫调 chénqiānglàndiào [clichéd]

(D) 老生常談 lǎoshēngchángtán [commonplace]

In Item 26, it is easy to directly understand the meaning of kǎnkǎnértán from the word itself, which means that, when conversing with Xiao-Quang, he is always "eloquent" and there is no limit to the topics in the conversation. Test-takers are immediately able to provide the correct answer, B duìdárúliú, which also has the same meaning as kǎnkǎnértán in Chinese. Therefore, according to the CEFR, Item 26 is categorised as being accounted for by an unidimensional ability, linguistics ability (Category I; corresponding to 5.2.1.1 of CEFR P.110) [46]. This item is implemented as a separated task item in this study.

Item 25 is an example of integrated task item categorization in Category III.

Item 25

小王：妳跟小李的關係是什麼？小李：泛泛之交。「泛泛之交」（fànfànzhījiāo）表示：

XiǎoWáng: Are you and XiǎoLǐ close friends or fànfànzhījiāo [cautious acquaintances]? What is the meaning behind the words, fànfànzhījiāo [cautious acquaintances]?

(A) 交情深厚 jiāoqíngshēnhòu [XiǎoWáng has been thick with XiǎoLǐ for years.]

(B) 交情一般 jiāoqíngyìbān [XiǎoWáng and XiǎoLǐ are casual friends.]

(C) 酒肉朋友 jiǔròupéngyǒu [XiǎoWáng and XiǎoLǐ are fair-weather friends.]

(D) 毫無交集 háowújiāojí [There is no interaction between XiǎoWáng and XiǎoLǐ.]

The meanings behind Item 25 are that XiǎoLǐ is a friend that XiǎoWáng knew considered as a common friend. According to the rhetoric forms used in Chinese conversation, option C (jiǔròupéngyǒu) conveys negative image. The stem of this item did not mention that XiǎoLǐ had bad influence on XiǎoWáng. Therefore, the correct answer of Item 25 is option B.

Item 25 is implemented not only to measure test-takers' linguistics ability in terms of their understanding of the Chinese proverb, fànfànzhījiāo [cautious acquaintances], but is also designed to evaluate test-takers' understanding of the meanings from the context and behind it. In addition, this item can be used to evaluate the mastery of the coherence of the context. Therefore, according to the CEFR, Item 25 is categorized as simultaneously being accounted for by linguistic, sociolinguistic and pragmatic competences (Category III; corresponding to 5.2.1.1 on p.110, and to 5.2.2.3 on p.120, and 5.2.3.2 on p.125 of the CEFR, respectively) [46].

Linguistic, sociolinguistic, and pragmatic ability each had 24 items, and there was a total of 43 items. Figure 2 shows the item structure of this study, which, for example, indicates that seven items are accounted for by the combined category of linguistic, sociolinguistic, and pragmatic ability (Category III). Those seven integrated task items are represented by Items 8, 9, 11, 15, 22, 23, and 25. The framework of the this study was developed and constructed so that it consisted of one overall and three domain abilities in a hierarchical ability structure, as shown in Figure 3.

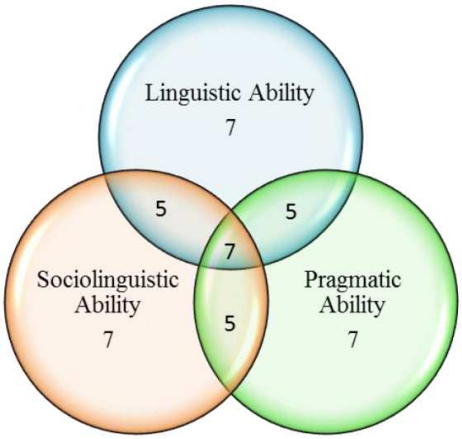

FIGURE 2. Item structure

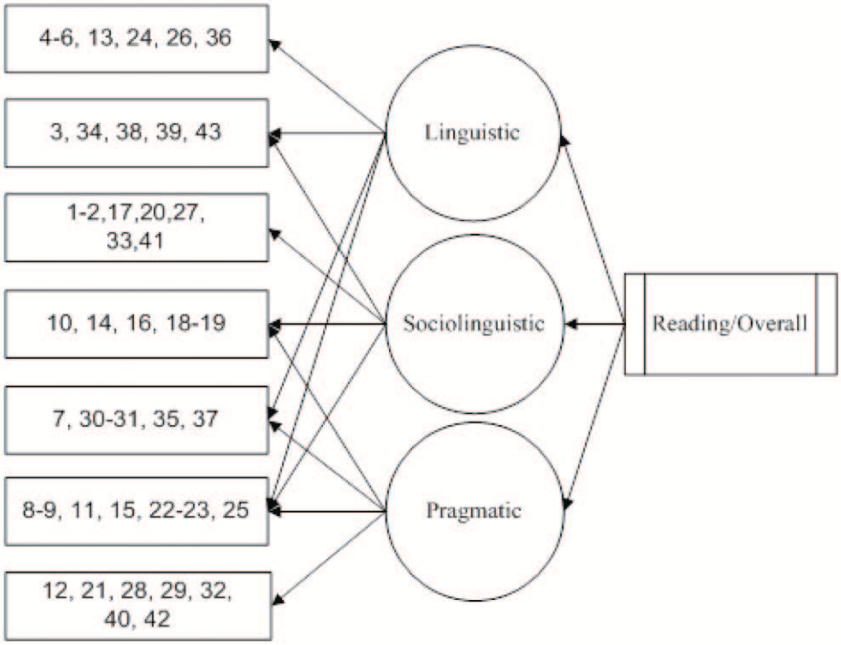

FIGURE 3. Hierarchical ability structure

3.2. **Ability estimation procedure.**

3.2.1. *Test structure of HCAT system.* The multidimensional three parameter logistic model (M3PLM) assumes that the test-taker's abilities and item discrimination are under a k dimensional vector, as for every response on the j item. The probability of i^{th} test-taker with ability $\boldsymbol{\Theta_i}$ correctly responding to item j will be concurrently influenced by k domain abilities (k dimensional vector), k discrimination parameters (k dimensional vector), a scalar of j^{th} item difficulty parameter, and a scalar of j^{th} item guessing parameter. Therefore, to enable the scalar of item difficulty parameter and the k dimensional ability vector to be subtracted, the scalar of item difficulty parameter b_j is multiplied by a $k \times 1$ unit matrix I indicated in Equation (1).

$$P_j(\Theta_i) = c_j + (1 - c_j)\frac{\exp[a_j'(\Theta_i - b_j I)]}{1 + \exp[a_j'(\Theta_i - b_j I)]} \quad i = 1, 2, \ldots, N; \;\; j = 1, 2, \ldots, n \qquad (1)$$

where $\boldsymbol{a}'_j = (a_{j1}, a_{j2}, \ldots, a_{jk})$ represents j^{th} item's k dimensional discrimination parameter vector; b_j is the scalar of difficulty parameter; c_j is the scalar of guessing parameter; and $\boldsymbol{\Theta}_i = (\theta_i^{(1)}, \theta_i^{(2)}, \ldots, \theta_i^{(k)})'$ represents i^{th} test-taker's k dimensional ability vector.

If the one factor within-item HO-IRT structure excludes the overall ability, it is considered to be a special case of the within-item MIRT structure. In other words, the test framework of the one-factor within-item HO-IRT structure assumes that some items are simultaneously accounted for by two or more domain abilities, whereas the correlation among the domain abilities is accounted for by positing a higher-order ability θ_i, which can be viewed as being the test-taker's overall ability, $\theta_i \sim N(0, 1)$. The i test-taker's k^{th} domain ability, $\theta^{(k)}$, is expressed as a linear function of the overall ability as shown in Equation (2).

$$\theta_i^{(k)} = \lambda^{(k)}\theta_i + \varepsilon_{ik} \tag{2}$$

where $\lambda^{(k)}$ is the latent coefficient when regressing the k domain ability on the overall ability; ε_{ik} is the error term, assuming a normal distribution with a mean of zero and variance of $1 - \lambda^{(k)2}$. The marginal distribution of each ability domain is also followed by the standard normal distribution (i.e., $\theta_i^{(k)} \sim N(0, 1)$).

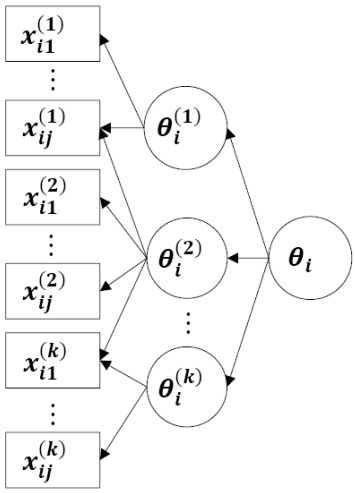

FIGURE 4. Diagram of the one-factor within-item HO-IRT structure

A diagram of the one-factor within-item HO-IRT structure is presented in Figure 4. For example, $x_{ij}^{(1)}$ is the j item responded to by the i test-taker and is also simultaneously accounted for by $\theta_i^{(1)}$ and $\theta_i^{(2)}$. Thus, $x_{ij}^{(1)}$ represents an integrated task item, which was implemented to measure two distinctive domain abilities at the same time. Therefore, the HO-IRT three parameter logistic model (H3PLM) of a one-factor within-item HO-IRT structure was able to simultaneously calibrate each individual domain ability together with the overall ability within the same framework and the test-taker's scores can be compared because all the domain and overall abilities are on the same scale [38,40,42]. The mathematical formula derivation undertook for abilities estimation was indicated as bellowed [44,47].

3.2.2. *Maximum likelihood estimation estimator.* Throughout the one-factor within-item HO-IRT structure in the HCAT system, a set of observed binary responses, $\boldsymbol{X} = (X_1, X_2, \ldots, X_N)$, from a single test-taker is assumed to possess local independence. Therefore, the likelihood function can be obtained by multiplying the probability functions for the

test-taker according to this observed response set. The likelihood function is shown as Equation (3):

$$L(\boldsymbol{X}|\boldsymbol{\theta}) = L(X_1, X_2, \ldots, X_N|\boldsymbol{\theta}) = \Pi_{j=1}^{N} P_j^{X_j}(\boldsymbol{\theta}) Q_j^{1-X_j}(\boldsymbol{\theta}) \tag{3}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)'$ indicates a K dimensional domain ability vector; $P_j(\boldsymbol{\theta})$ is the probability function defined in Equation (1). $Q_j(\boldsymbol{\theta}) = 1 - P_j(\boldsymbol{\theta})$.

Firstly, to obtain the log-likelihood function, $\ln L(\boldsymbol{X}|\boldsymbol{\theta})$, the maximum likelihood estimation (MLE) estimator is the solution to estimate the K domain abilities, which can maximize the first derivatives of $\ln L(\boldsymbol{X}|\boldsymbol{\theta})$ given by Equation (4):

$$\frac{\partial}{\partial \theta} \ln L(\boldsymbol{X}|\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial}{\partial \theta_1} \ln L(X_1, X_2, \ldots, X_N|\boldsymbol{\theta}) \\ \frac{\partial}{\partial \theta_2} \ln L(X_1, X_2, \ldots, X_N|\boldsymbol{\theta}) \\ \vdots \\ \frac{\partial}{\partial \theta_K} \ln L(X_1, X_2, \ldots, X_N|\boldsymbol{\theta}) \end{bmatrix} \tag{4}$$

where

$$\frac{\partial}{\partial \theta} \ln L(\boldsymbol{X}|\boldsymbol{\theta}) = \sum\nolimits_{j \in \boldsymbol{S}} v_j a_j \tag{5}$$

$$v_j = \frac{[P_j(\boldsymbol{\theta}) - c_j][X_j - P_j(\boldsymbol{\theta})]}{(1 - c_j) P_j(\boldsymbol{\theta})} \tag{6}$$

$\boldsymbol{S}$ is a vector space that contains the items already administered.

Secondly, since Equation (4) has no closed form solution, the iterative Newton-Raphson procedure can calibrate an approximation that maximizes Equation (4). The approximation function can be written as Equation (7):

$$\boldsymbol{\theta}^{(m)} = \boldsymbol{\theta}^{(m-1)} - \boldsymbol{\delta}^{(m)} \tag{7}$$

where $\boldsymbol{\theta}^{(m)}$ is the m approximation to the value of $\boldsymbol{\theta}$

$$\boldsymbol{\delta}^{(m)} = [H(\boldsymbol{\theta}^{(m)})]^{-1} \times \left[\frac{\partial \ln L(\boldsymbol{X}|\boldsymbol{\theta}^{(m)})}{\partial \boldsymbol{\theta}} \right] \tag{8}$$

$H(\boldsymbol{\theta}^{(m)})$ represents the m element of $H(\boldsymbol{\theta})$ which is a $K \times K$ symmetrix matrix of the second derivatives of $\ln L(\boldsymbol{X}|\boldsymbol{\theta})$ as indicated in Equation (9):

$$H(\boldsymbol{\theta}) = \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \ln L(\boldsymbol{X}|\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial^2}{\partial \theta_1^2} \ln L(\boldsymbol{X}|\boldsymbol{\theta}) \ldots\ldots\ldots\ldots\ldots \frac{\partial^2}{\partial \theta_1 \partial \theta_K} \ln L(\boldsymbol{X}|\boldsymbol{\theta}) \\ \frac{\partial^2}{\partial \theta_2^2} \ln L(\boldsymbol{X}|\boldsymbol{\theta}) \ldots\ldots \frac{\partial^2}{\partial \theta_2 \partial \theta_K} \ln L(\boldsymbol{X}|\boldsymbol{\theta}) \\ \vdots \\ \vdots \\ \frac{\partial^2}{\partial \theta_K^2} \ln L(\boldsymbol{X}|\boldsymbol{\theta}) \end{bmatrix} \tag{9}$$

$$H(\boldsymbol{\theta}) = \sum\nolimits_{j \in \boldsymbol{S}} a_j a_j' w_j; \quad w_j = \frac{[P_j(\boldsymbol{\theta}) - c_j][c_j X_j - P_j^2(\boldsymbol{\theta})] Q_j(\boldsymbol{\theta})}{(1 - c_j)^2 P_j^2(\boldsymbol{\theta})} \tag{10}$$

The diagonal element of (t, t) in $H(\boldsymbol{\theta})$ is shown in Equation (11):

$$\frac{\partial^2}{\partial \theta_t^2} \ln L(\boldsymbol{X}|\boldsymbol{\theta}) = \sum\nolimits_{j \in \boldsymbol{S}} \frac{a_{jt}^2 [P_j(\boldsymbol{\theta}) - c_j][c_j X_j - P_j^2(\boldsymbol{\theta})] Q_j(\boldsymbol{\theta})}{(1 - c_j)^2 P_j^2(\boldsymbol{\theta})} \tag{11}$$

The off-diagonal element of (t, u) in $H(\boldsymbol{\theta})$ is shown in Equation (12):

$$\frac{\partial^2}{\partial \theta_t \partial \theta_u} \ln L(\boldsymbol{X}|\boldsymbol{\theta}) = \sum\nolimits_{j \in \boldsymbol{S}} \frac{a_{jt} a_{ju} [P_j(\boldsymbol{\theta}) - c_j][c_j X_j - P_j^2(\boldsymbol{\theta})] Q_j(\boldsymbol{\theta})}{(1 - c_j)^2 P_j^2(\boldsymbol{\theta})} \tag{12}$$

The HCAT system will then select the next $m+1$ item based on $\boldsymbol{\theta}^{(m)}$ for the test-taker to respond. If Equations (7) and (8) are unable to converge, Equation (9) can be replaced by using Fisher's method of scoring, as shown in Equation (13):

$$I(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}) = -E[H(\boldsymbol{\theta})] = -E\left[\frac{\partial^2}{\partial\theta\partial\theta'}\ln L(\boldsymbol{X}|\boldsymbol{\theta})\right] = \sum\nolimits_{j\in\boldsymbol{S}}\Psi_j \tag{13}$$

where $\Psi_j = a_j a_j' w_j^*$; $w_j^* = \frac{Q_j(\boldsymbol{\theta})}{P_j(\boldsymbol{\theta})} \times \left[\frac{P_j(\boldsymbol{\theta})-c_j}{1-c_j}\right]^2$.

The diagonal element of (t,t) in $I(\boldsymbol{\theta},\hat{\boldsymbol{\theta}})$ is $I_{tt}(\boldsymbol{\theta},\hat{\boldsymbol{\theta}})$ shown in Equation (14):

$$I_{tt}(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}) = -\sum\nolimits_{j\in\boldsymbol{S}}\frac{a_{jt}^2[P_j(\boldsymbol{\theta})-c_j][c_jP_j(\boldsymbol{\theta})-P_j^2(\boldsymbol{\theta})]Q_j(\boldsymbol{\theta})}{(1-c_j)^2P_j^2(\boldsymbol{\theta})} \tag{14}$$

The off-diagonal element of (t,u) in $I(\boldsymbol{\theta},\hat{\boldsymbol{\theta}})$ is $I_{tu}(\boldsymbol{\theta},\hat{\boldsymbol{\theta}})$ shown in Equation (15):

$$I_{tu}(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}) = -\sum\nolimits_{j\in\boldsymbol{S}}\frac{a_{jt}a_{ju}[P_j(\boldsymbol{\theta})-c_j][c_jP_j(\boldsymbol{\theta})-P_j^2(\boldsymbol{\theta})]Q_j(\boldsymbol{\theta})}{(1-c_j)^2P_j^2(\boldsymbol{\theta})} \tag{15}$$

3.2.3. *Maximum a posteriori estimator.* Lord [48] and Mislevy [49] adopted test-takers' posterior distribution and introduced a maximum a posteriori (MAP) estimator to measure their abilities. The MAP estimator is the maximizer of the posterior density function of $\boldsymbol{\theta}$. The posterior density function of $\boldsymbol{\theta}$ is formed by being weighted from test-takers' prior distribution function. It is defined as $f(\boldsymbol{\theta}|\boldsymbol{X})$ and shown in Equation (16).

$$f(\boldsymbol{\theta}|\boldsymbol{X}) = L(\boldsymbol{X}|\boldsymbol{\theta})\frac{f(\boldsymbol{\theta})}{f(\boldsymbol{X})},\ \ L(\boldsymbol{X}|\boldsymbol{\theta}) \text{ was defined in Equation (3)} \tag{16}$$

The prior density function, $f(\boldsymbol{\theta})$, can be written as Equation (17) and $f(\boldsymbol{X})$ is the marginal probability of $\boldsymbol{X}$:

$$f(\boldsymbol{\theta}) = (2\pi)^{-D/2}|\boldsymbol{\Phi}|^{-1/2}\exp\left[-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu})'\boldsymbol{\Phi}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})\right] \tag{17}$$

Since the prior information consists of more than two abilities, the prior distribution was assumed as a multivariate normal distribution ($\boldsymbol{\theta}\sim MVN(\boldsymbol{\mu},\boldsymbol{\Phi})$) with a mean vector $\boldsymbol{\mu}$ and variance-and-covariance matrix $\boldsymbol{\Phi}$. The same procedures were adopted as in the MLE estimator. Firstly, the log-likelihood function, $\ln f(\boldsymbol{\theta}|\boldsymbol{X})$, was obtained as shown in Equation (18):

$$\begin{aligned}\ln f(\boldsymbol{\theta}|\boldsymbol{X}) &= \ln L(\boldsymbol{X}|\boldsymbol{\theta}) + \ln f(\boldsymbol{\theta}) + k\\ &= \ln L(\boldsymbol{X}|\boldsymbol{\theta}) - \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu})'\boldsymbol{\Phi}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu}) + c,\ \ c \text{ and } k \text{ are constant}\end{aligned} \tag{18}$$

The first derivatives of $\ln f(\boldsymbol{\theta}|\boldsymbol{X})$ can be written as Equation (19):

$$\frac{\partial}{\partial\boldsymbol{\theta}}\ln f(\boldsymbol{\theta}|\boldsymbol{X}) = \frac{\partial}{\partial\boldsymbol{\theta}}\ln L(\boldsymbol{X}|\boldsymbol{\theta}) - \frac{1}{2}\frac{\partial}{\partial\boldsymbol{\theta}}[(\boldsymbol{\theta}-\boldsymbol{\mu})'\boldsymbol{\Phi}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})] \tag{19}$$

where

$$\frac{\partial}{\partial\boldsymbol{\theta}}\ln f(\boldsymbol{\theta}|\boldsymbol{X}) = \sum\nolimits_{j\in\boldsymbol{S}}v_ja_j - \boldsymbol{\Phi}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu}) \tag{20}$$

$$\frac{\partial}{\partial\theta_t}\ln f(\boldsymbol{\theta}|\boldsymbol{X}) = \sum\nolimits_{j\in\boldsymbol{S}}\frac{a_{jt}[P_j(\boldsymbol{\theta})-c_j][X_j-P_j(\boldsymbol{\theta})]}{(1-c_j)P_j(\boldsymbol{\theta})} - \left[\frac{\partial}{\partial\theta_t}(\boldsymbol{\theta}-\boldsymbol{\mu})'\right]\boldsymbol{\Phi}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu}),\ \ t=1,2,\ldots,K \tag{21}$$

Secondly, the iterative New-Raphson procedure in MAP estimator can be implemented in Equation (22).

$$\boldsymbol{\theta}^{(j)} = \boldsymbol{\theta}^{(j-1)} - \boldsymbol{\delta}^{(j)} \tag{22}$$

where

$$\boldsymbol{\delta}^{(j)} = [M(\boldsymbol{\theta}^{(j)})]^{-1} \times \left[\frac{\partial \ln f(\boldsymbol{\theta}^{(j)}|\boldsymbol{X})}{\partial \boldsymbol{\theta}}\right] \tag{23}$$

$$M(\boldsymbol{\theta}) = \frac{\partial^2}{\partial\theta\partial\theta'} \ln L(\boldsymbol{\theta}|\boldsymbol{X}) = \begin{bmatrix} \frac{\partial^2}{\partial\theta_1^2} \ln f(\boldsymbol{\theta}|\boldsymbol{X}) \ldots\ldots\ldots\ldots\ldots \frac{\partial^2}{\partial\theta_1\partial\theta_K} \ln f(\boldsymbol{\theta}|\boldsymbol{X}) \\ \frac{\partial^2}{\partial\theta_2^2} \ln f(\boldsymbol{\theta}|\boldsymbol{X}) \ldots\ldots \frac{\partial^2}{\partial\theta_2\partial\theta_K} \ln f(\boldsymbol{\theta}|\boldsymbol{X}) \\ \vdots \\ \vdots \\ \frac{\partial^2}{\partial\theta_K^2} \ln f(\boldsymbol{\theta}|\boldsymbol{X}) \end{bmatrix} \tag{24}$$

$$M(\boldsymbol{\theta}) = \sum\nolimits_{j\in\boldsymbol{S}} a_j a_j' w_j - \Phi^{-1}, \;\; w_j \text{ is same as Equation (10)} \tag{25}$$

The diagonal element of (t,t) in $M(\boldsymbol{\theta})$ is shown in Equation (26):

$$\frac{\partial^2}{\partial\theta_t^2} \ln f(\boldsymbol{\theta}|\boldsymbol{X}) = \sum\nolimits_{j\in\boldsymbol{S}} \frac{a_{jt}^2[P_j(\boldsymbol{\theta}) - c_j][c_jX_j - P_j^2(\boldsymbol{\theta})]Q_j(\boldsymbol{\theta})}{(1-c_j)^2P_j^2(\boldsymbol{\theta})} - \Phi^{tt} \tag{26}$$

The off-diagonal element of (t,u) in $M(\boldsymbol{\theta})$ is shown in Equation (27):

$$\frac{\partial^2}{\partial\theta_t\partial\theta_u} \ln f(\boldsymbol{\theta}|\boldsymbol{X}) = \sum\nolimits_{j\in\boldsymbol{S}} \frac{a_{jt}a_{ju}[P_j(\boldsymbol{\theta}) - c_j][c_jX_j - P_j^2(\boldsymbol{\theta})]Q_j(\boldsymbol{\theta})}{(1-c_j)^2P_j^2(\boldsymbol{\theta})} - \Phi^{tu} \tag{27}$$

Φ^{tt} and Φ^{tu} represent the t diagonal and the (t,u) elements of $\boldsymbol{\Phi}$, respectively.

3.2.4. *Overall ability estimated.* Since each of domain ability in the one factor within-item HO-IRT structure is expressed as a linear function of the overall ability, the overall ability can be estimated through a linear transformation from these domain abilities. The correlation between the domain abilities is shown as Equation (28).

$$\mathbb{R} = \begin{bmatrix} 1 & \lambda^{(1)}\lambda^{(2)} & \ldots & \lambda^{(1)}\lambda^{(k)} \\ & 1 & \ldots & \lambda^{(1)}\lambda^{(k)} \\ & & \ddots & \vdots \\ & & & 1 \end{bmatrix} \tag{28}$$

where $\mathbb{R}$ is a symmetric matrix and can be adopted as the test-taker's prior information.

The single scalar of overall ability is estimated by Equation (29).

$$\theta_H = \boldsymbol{\lambda}'\mathbb{R}^{-1}\boldsymbol{\theta_L} \tag{29}$$

θ_H and $\boldsymbol{\theta_L}$ represent the scalar of overall ability and the vector of domain ability, respectively. $\boldsymbol{\lambda}' = [\lambda^{(1)}, \lambda^{(2)}, \ldots, \lambda^{(k)}]$.

3.3. **Item selection procedure.**

3.3.1. *Item selection procedure of the MLE estimator.* The mathematical formula derivation undertook for item selection procedure was indicated as bellowed [44,47]. According to Equations (14) and (15), each element of $I(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ is formed from item level summands. Therefore, $I(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ based on an item information can be denoted by $I(\boldsymbol{\theta}, X_j)$ where the diagonal element of (t,t) in $I(\boldsymbol{\theta}, X_j)$ is shown in Equation (30):

$$I_{tt}(\boldsymbol{\theta}, X_j) = \frac{\left[\frac{\partial P_j(\boldsymbol{\theta})}{\partial\theta_t}\right]^2}{P_j(\boldsymbol{\theta}) \cdot Q_j(\boldsymbol{\theta})} \tag{30}$$

The off-diagonal element of (t, u) in $I(\boldsymbol{\theta}, X_j)$ is shown in Equation (31):

$$I_{tu}(\boldsymbol{\theta}, X_j) = \frac{\frac{\partial P_j(\boldsymbol{\theta})}{\partial \theta_t} \times \frac{\partial P_j(\boldsymbol{\theta})}{\partial \theta_u}}{P_j(\boldsymbol{\theta}) \cdot Q_j(\boldsymbol{\theta})} \tag{31}$$

The provisional ability, $\hat{\boldsymbol{\theta}}_{\boldsymbol{S}}$, and item information, $I(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{\boldsymbol{S}})$, are calibrated and obtained after the test-taker responds to the first $m-1$ item. The main idea of the item selection procedure of the MLE estimator is to find the item that maximizes Equation (32).

$$|I(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{\boldsymbol{S}}) + I(\boldsymbol{\theta}, \boldsymbol{X}_{\boldsymbol{S}'})| \tag{32}$$

where $\boldsymbol{S}$ and $\boldsymbol{S}'$ are the vector space the items administered and those that remained, respectively.

3.3.2. *Item selection procedure of the MAP estimator.* The difference between the MAP and the MLE estimator is that the former incorporates a posterior probability density function on the item selection information whereas the latter uses a log-likelihood function. If $\boldsymbol{\Phi}^{-1}$ is excluded from Equation (25), the two equations, (25) and (10), are identical. Therefore, the item information of the MAP estimator can be carried out by adding $\boldsymbol{\Phi}^{-1}$ to Equation (32). The item selection procedure of the MAP estimator in the MCAT system is indicated as Equation (33).

$$|I(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{\boldsymbol{S}}) + I(\boldsymbol{\theta}, \boldsymbol{X}_{\boldsymbol{S}'}) + \boldsymbol{\Phi}^{-1}| \tag{33}$$

3.4. **HCAT system development.** The HCAT system was developed as an internet-based computerized testing system with a three-tier client/server architecture, which enabled the system to be operated from both the client-side and the server-side. The client-side was designed as the user interface, and the test-takers could log into the system remotely through a web browser via HTML. Please refer to Figure 5.

The server-side was operated via CentOS 5 (Linux version). MySQL and Apache were applied as the data storage and Web server software respectively. The architecture of the HCAT system is shown in Figure 6.

FIGURE 5. Login interface in client-side

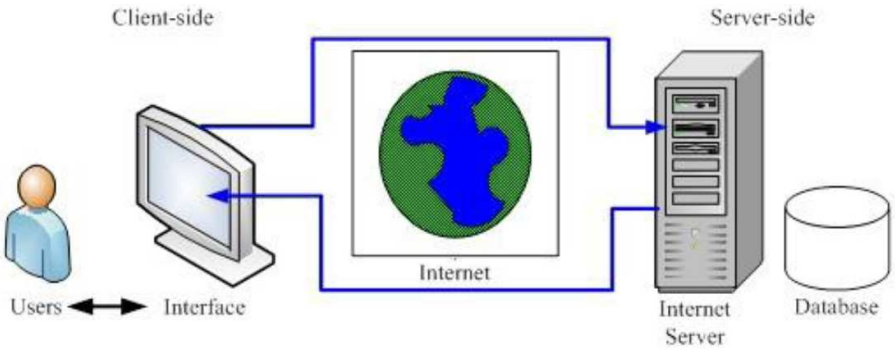

FIGURE 6. The HCAT system architecture

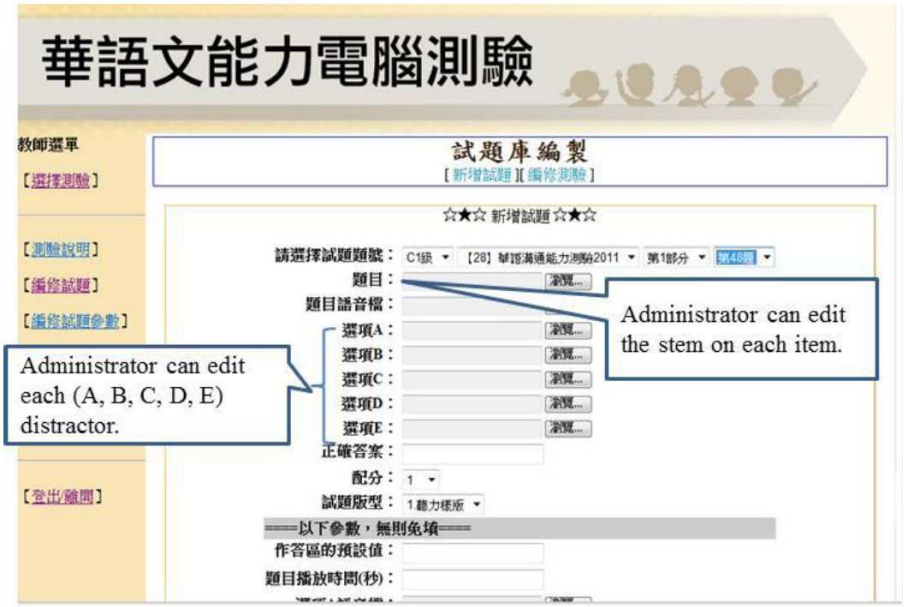

FIGURE 7. Item bank management interface

3.4.1. *Item bank.* The server-side of the HCAT system has four different management functions, which are divided into item bank, test assignment, response, and proficiency report. The system administrator is permitted to log on to the system to edit items through an item bank management interface, as shown in Figure 7.

3.4.2. *Test assignment.* The system administrator can assign any particular test to the test-taker through the test assignment management interface, as shown in Figure 8. In the test process type selection for "IRT adaptive testing", the HCAT system will select the next item for test-takers to respond to according to the items to which they have already responded. The H3PLM is applied in this study, and three item parameters can be calibrated. In Figure 9, the item parameter editing function allows the system administrator to input the scale of item parameters.

3.4.3. *Response.* Figure 10 shows the response management interface, with the current item test number in the top left-hand corner and the time remaining for the test in the top right-hand corner. The stem of the item is on the bottom left-hand side and the options for this item are shown on the bottom right-hand side. Different formats of item stem or options are enabled in the HCAT system for different screen display formats. Other than the text format, the item stem or the options can be displayed with pictures or a video.

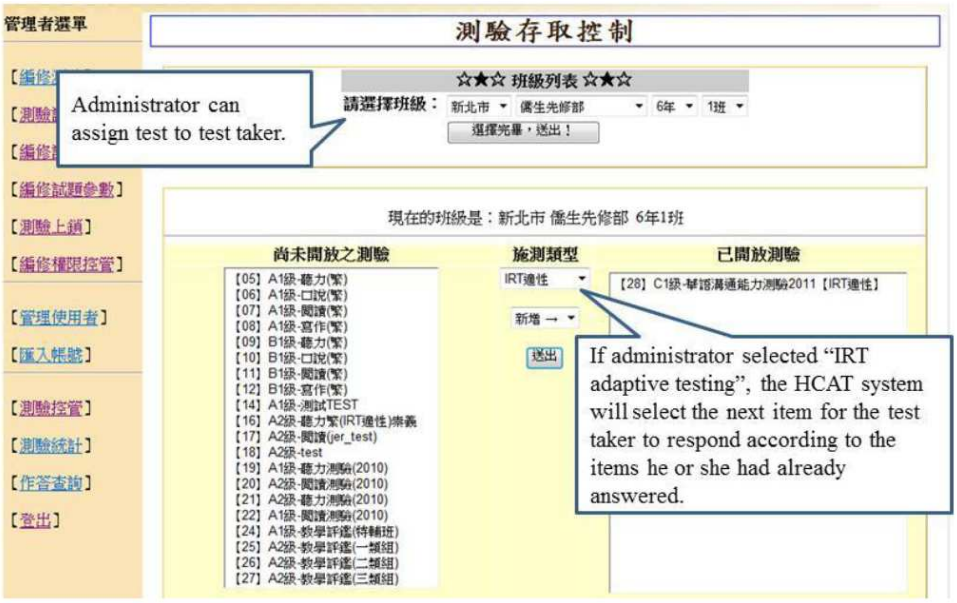

FIGURE 8. Test assignment management interface (1)

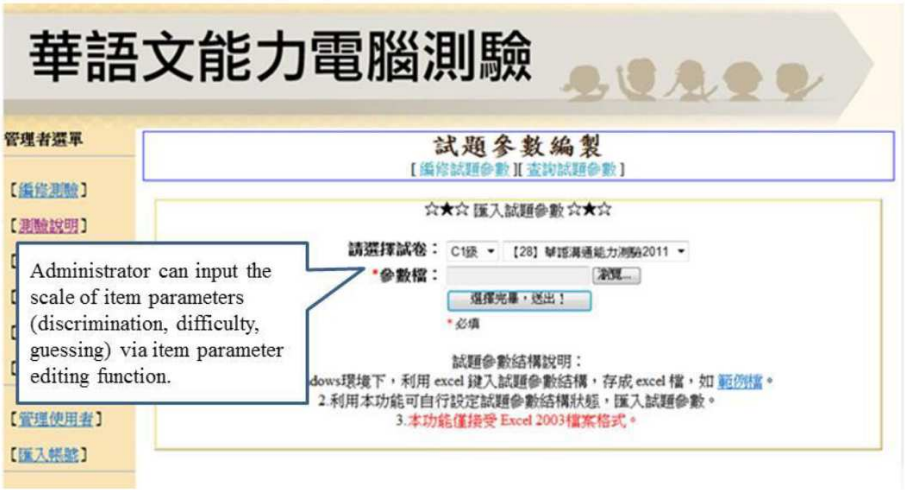

FIGURE 9. Test assignment management interface (2)

Test-takers can click on the correct option on the screen to complete their response to this item. The screen will retain the same item until the test-taker clicks the "next" button to continue to the next item.

3.4.4. *Proficiency report.* Having completed the test, test-takers are able to check their results through the proficiency report interface, which provides a report download function. The report includes the test date and the time taken to respond to each item, as shown in Figure 11.

4. **Data Analysis.** The empirical data was derived from a computer-based test in April 2011, which was undertaken at the National Taiwan Normal University. A total number of 1,272 test-takers took the test, and the results of 1,235 of them were used for the analysis. Each test-taker was examined in 43 individual items, and any pattern of missing data was excluded.

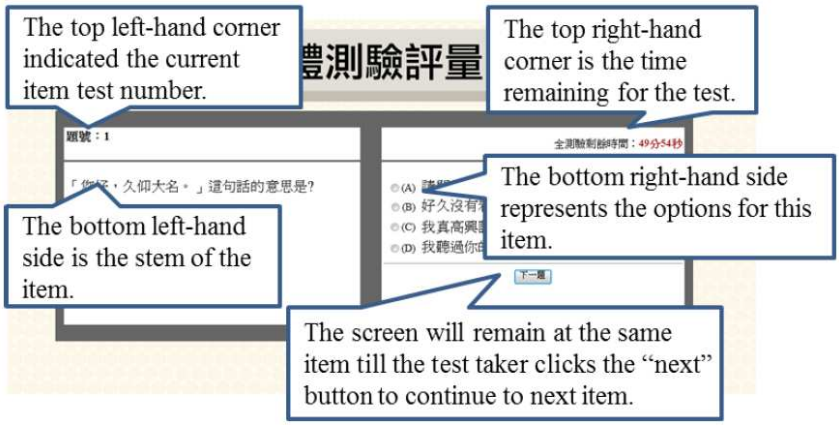

FIGURE 10. Response management interface

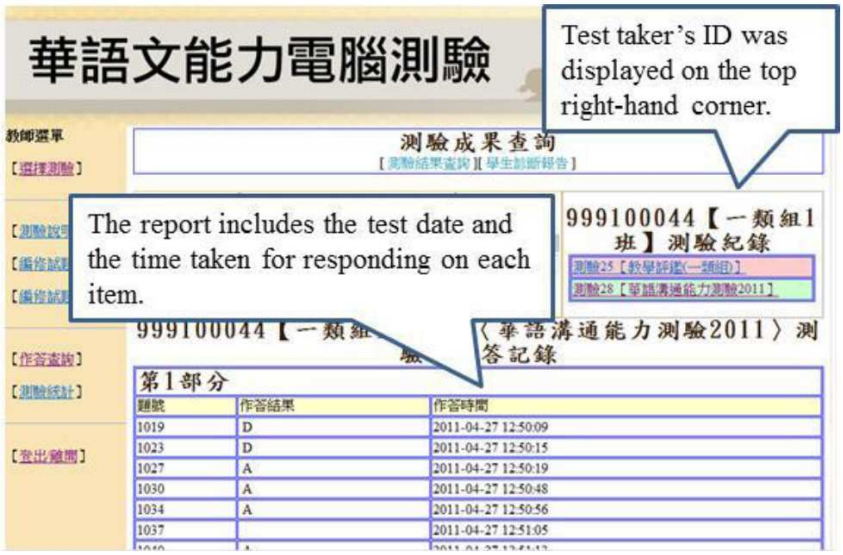

FIGURE 11. Proficiency report interface

TABLE 1. Correlation analysis of three domain abilities

	Linguistic	Sociolinguistic	Pragmatic
Linguistic	1	0.7017	0.6817
Sociolinguistic	0.7017	1	0.8407
Pragmatic	0.6817	0.8407	1

4.1. **Three domain and overall abilities estimation.** The item parameters and the test-takers' three domain and overall abilities were calibrated in WinBUGS with the Markov Chain Monte Carlo (MCMC) method.

According to Table 1, the correlation coefficient was 0.7017, 0.6817 and 0.8407 for linguistic and sociolinguistic ability, linguistic and pragmatic ability, and sociolinguistic and pragmatic ability, respectively. According to Table 2, the average of linguistic, sociolinguistic, and pragmatic ability was 0.0168, 0.0186 and 0.0176, respectively. The standard errors of these three domain abilities were 0.0220, 0.0230 and 0.0231, respectively. In addition, the latent coefficients when regressing the linguistic, sociolinguistic and pragmatic

TABLE 2. Three domain and overall abilities estimation

	Mean	S.E	coefficient
Linguistic	0.0168	0.0220	0.7631
Sociolinguistic	0.0186	0.0230	0.9348
Pragmatic	0.0176	0.0231	0.9059
Overall	0.0188	0.0232	

ability on the overall ability were 0.7631, 0.9348, and 0.9059, respectively. The average of the test-takers' overall ability and standard error was 0.0188 and 0.0232, respectively.

4.2. **Simulation study on HCAT system.** All the item parameters, as well as the test-takers' overall and domain abilities, were calibrated on the basis of the item responses. These abilities will be used as the true value in this simulation study. The HCAT system will repeatedly calibrate test-takers' provisional ability according to their response to the item that is simulated and assigned by the item selection procedure in the HCAT system. The root mean square of error ($RMSE$) was adopted as the evaluation criterion to compare the accuracy and efficiency of the ability estimation and item selection procedure of the MLE and the MAP estimator.

4.2.1. *Criterion of domain ability evaluation.* The $RMSE$ of the domain ability evaluation criterion is described below in Equation (34):

$$RMSE(L) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\theta}_i^{(K)} - \theta_i^{(K)}\right)^2} \tag{34}$$

where $\theta_i^{(K)}$ represents the i test-taker's k^{th} domain ability, and $\hat{\theta}_i^{(K)}$ represents the estimation of this particular test-taker's k^{th} domain ability.

The accuracy of the three domain abilities estimation of the MLE and MAP estimator had shown in Figure 12 and Figure 13, respectively.

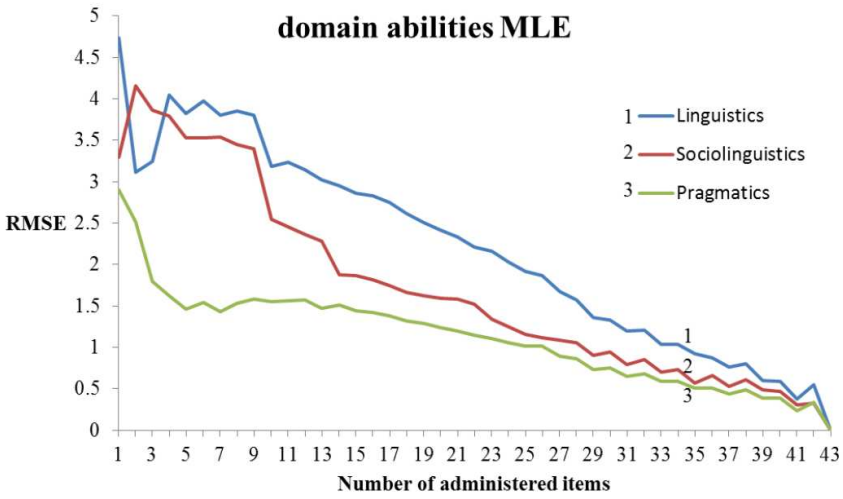

FIGURE 12. The accuracy of domain abilities estimation of the MLE estimator

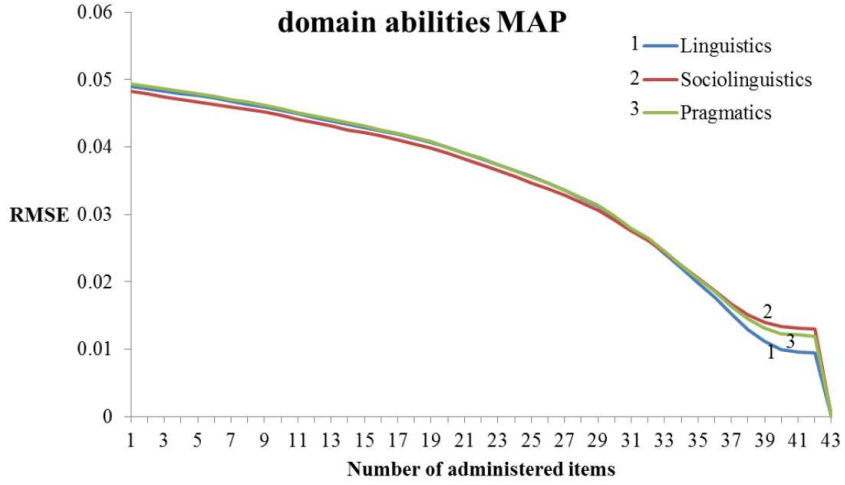

FIGURE 13. The accuracy of domain abilities estimation of the MAP estimator

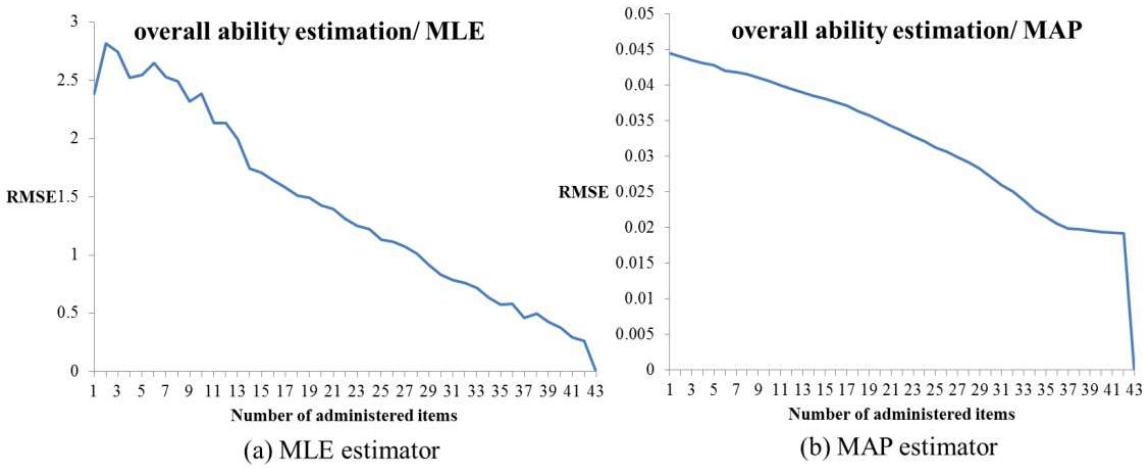

FIGURE 14. Overall ability estimation comparison between the MLE and MAP estimator: (a) MLE; (b) MAP

4.2.2. *Criterion of overall ability evaluation.* The $RMSE$ of the overall ability evaluation criterion is described below in Equation (35):

$$RMSE(H) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\theta}_i - \theta_i\right)^2} \tag{35}$$

where θ_i represents the i test-taker's overall ability, $\hat{\theta}_i$ represents the estimation of this particular test-taker's overall ability, and N is the total numbers of test-takers.

The accuracy of the ability estimation for overall ability comparison between the MLE and MAP estimator had shown in Figure 14.

The efficiency of the item selection procedure evaluation for the MAP estimator had shown in Figure 15.

5. **Discussion and Conclusions.** The Fitness comparison for the testing structure, the AIC, BIC and DIC statistics of H3PLM (47334, 67400 and 121319, respectively) were smaller compared with those of M3PLM (48123, 68188 and 123147, respectively). These empirical data fits well with the one factor within-item HOIRT structure compared to the within-item MIRT structure. In addition, according to Table 1, the three domain

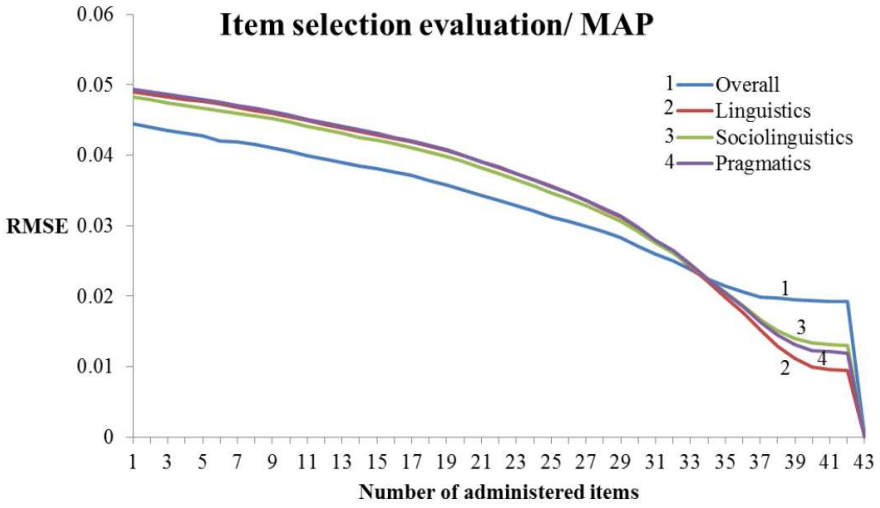

FIGURE 15. Item selection procedure evaluation for the MAP estimator

abilities of reading ability in Chinese are correlated. According to Table 2, the latent coefficients when regressing the linguistic, sociolinguistic and pragmatic ability on the overall ability are relative high. These results provides further evidence in relative to other researches via CFA method on dichotomous data analysis that CLA in Chinese is multi-componential and consists of a general higher-order factor. The application of the one factor within-item HO-IRT structure is suitable to adopt in the HCAT system.

The procedures to estimate ability and select items in the HCAT system are much more complex. For example, by comparing the traditional paper-and-pencil test or computerized test with the HCAT system on ability estimation and item selection procedure, the HCAT system needs to continuously calibrate test-takers' ability from their current response and select the next item that is able to maximize item information, which is indicated in Equation (33). The standard error of estimation is equivalent to the reciprocal of the square root of the amount of item information; the larger the item information calibrated, the smaller the standard error of estimation. According to Equation (33), the item selection procedure of the MAP estimator comprises three components: an inverse of variance-and-covariance matrix and two item information matrices. Therefore, if the next item selected for the test-taker to respond to is still accounted for by the same category as the previously-administrated item, then the mechanism of maximized Equation (33) will be violated. In addition, this study adopted empirical data and multidimensional item parameters to conduct a simulated study of the HCAT system, which was developed and implemented based on one factor within-item HO-IRT structure. According to Figure 15, the $RMSE$ values of the MAP estimator for the three domain abilities were similar, and they were all smaller than the $RMSE$ value of overall ability when the test-taker had completed and responded to 34 items. The results indicate that the estimation errors in the domain abilities were accumulated and contributed to the overall ability, and this verified the appropriateness of applying this item selection procedure in the HCAT system. Therefore, the efficiency of the mathematical formula derivation undertook for item selection procedure is able to ensure.

Regards to the comparison (Figures 12 and 13) on the accuracy of the domain ability estimation of the MLE and MAP estimators in the HCAT system, the $RMSE$ curves of the MLE and MAP estimators were both decreased when the numbers of items administered to the test-takers were progressively increased. This result indicated that both the MLE and MAP estimators were able to accurately calibrate the test-takers' three domain

abilities. However, the ability estimation of the MAP estimator was more accurate than that of the MLE estimator, and this was because the curves of the MAP estimator were smoother and descended steadily compared with those of the MLE estimator. Secondly, in terms of comparing the efficiency of the item selection procedure of the MLE and MAP estimators in the HCAT system, according to Figures 14(a) and 14(b), the results on the accuracy of the overall ability estimation indicated that the MAP estimator was more accurately than the MLE estimator. This was because the $RMSE$ value of overall ability was below 0.04 as soon as the test-takers had responded to 10 of the administered items to which the MAP estimator was applied to the HCAT system. However, when the MLE estimator was applied in the HCAT system, the $RMSE$ value was inconsistent with the numbers of items responded to and more items were required to be completed in order to reach a $RMSE$ below 0.04. Thus, the HCAT system adopted the MAP estimator administered about one-third of the total items than it adopted the MLE estimator. In addition, the application of the MAP estimator enables as more accurately calibrating test-takers' three domain abilities and reading ability in Chinese.

The developed the HCAT system enables the test-taker to participate 10 of 43 items, while their three domain abilities can be measured and their overall ability can be directly reported. This makes the HCAT system very handy and practical. First of all, the test administration time and costs can be reduced. Secondly, the results make CPT a multi-functional and practical test. For example, the linguistic, sociolinguistic and pragmatic ability estimated from the HCAT system can be used as a formative assessment for diagnostic purposes, and the overall ability can be provided the overall level of performance for a summative assessment.

This study has succeeded in making four major contributions. Firstly, the HCAT system is the first computerized adaptive test system originally developed for the CPT. Secondly, the application of one factor within-item HO-IRT structure enables the item selection procedure for selecting the next item for the test-taker to respond to is simultaneously based on the test-taker's three domain abilities, while some of the items are accounted for by two or more abilities. Thirdly, the application of the MAP estimator enables the test-takers to respond to fewer items as well as their abilities can be accurately calibrated. Last, but not least, the HCAT system is able to directly reduce the test administration costs compared with either the current traditional paper-and-pencil or computer-based CPTs.

Acknowledgement. The authors gratefully acknowledge the support of this study by the National Science Council of Taiwan, under the Grant No NSC 99-2410-H-142-008-MY3.

REFERENCES

[1] J. B. Carroll, The psychology of language testing, in *Language Testing Symposium. A Psycholinguistic Perspective*, A. Davies (ed.), London, Oxford University Press, 1968.
[2] R. Lado, *Language Testing*, McGraw-Hill, New York, 1961.
[3] L. F. Bachman and A. S. Palmer, *Language Testing in Practice*, Oxford University Press, London, 1997.
[4] Y. Sawaki, L. J. Sticker and A. Oranje, Factor structure of the TOEFL Internet-based test (iBT): Exploration in a field trial sample, *Educational Testing Services RR-04-24*, 2008.
[5] Y. Sawaki, L. J. Sticker and A. Oranje, Factor structure of the TOEFL Internet-based test, *Language Testing*, vol.26, pp.5-30, 2009.
[6] L. F. Bachman, F. Davidson, K. Ryan and I.-C. Choi, *An Investigation into the Comparability of Two Tests of English as a Foreign Language: The Cambridge-TOEFL Comparability Study*, Cambridge University Press, New York, 1995.

[7] K. A. Fouly, L. F. Bachman and G. A. Cziko, The divisibility of language competence: A confirmatory approach, *Language Learning*, vol.40, pp.1-21, 1990.
[8] J. W. Oller, *Language Tests at School*, Longman Group Ltd., London, 1979.
[9] J. W. Oller, *Issues in Language Testing Research*, Newbury House Publishers, Inc., Rowley, MA, 1983.
[10] J. W. Oller and J. Jonz, *Cloze and Coherence*, Associated University Press, London, 1994.
[11] C. A. Stone and C.-C. Yeh, Assessing the dimensionality and factor structure of multiple-choice exams: An empirical comparison of methods using the multistate bar examination, *Educational and Psychological Measurement*, vol.66, pp.193-214, 2006.
[12] T. A. Ackerman, M. J. Gierl and C. Walker, An NCME instructional module on using multidimensional item response theory to evaluate educational and psychological tests, *Educational Measurement: Issues and Practice*, vol.22, pp.37-53, 2003.
[13] S. B. Green, Identifiability of spurious factors with linear factor analysis with binary items, *Applied Psychological Measurement*, vol.7, 1983.
[14] K. A. Swygert, L. D. McLeod and D. Thissen, *Factor Analysis for Items or Testlets Scored in More than Two Categories*, Chapel Hill, University of North Carolina, L. L. Thurs Tone Psychometric Laboratory, 2001.
[15] R. D. Bock, R. Gibbons and E. J. Muraki, Full information item factor analysis, *Applied Psychological Measurement*, vol.12, pp.261-280, 1988.
[16] J. B. Carroll, The effect of difficulty and chance success on correlations between items or between tests, *Psychometrika*, vol.10, pp.1-19, 1945.
[17] M. Chalhoub-Deville, Theoretical models, assessment frameworks and test construction, *Language Testing*, vol.14, pp.3-22, 1997.
[18] P. Skehan, State-of-the-art article: Language testing, Part 1, *Language Testing*, vol.24, pp.211-221, 1988.
[19] D. Douglas, *Assessing Language for Specific Purposes*, Cambridge University Press, Cambridge, 2000.
[20] B. O'Sullivan and C. J. Weir, Test development and validation, in *Language Testing: Theories and Practices*, B. O'Sullivan (ed.), Palgrave Macmillan, UK, 2011.
[21] M. K. Wolf, J. Kao, J. Herman, L. F. Bachman, A. Bailey, P. L. Bachman, T. Farnsworth and S. M. Chang, Issues in assessing English language learners: English language proficiency measures and accommodation uses – Literature review (Part 1 of 3), *CRESST/UCLA, Los Angeles, CA CRESST Report No. 731*, 2008.
[22] L. Gu, At the interface between language testing and second language acquisition: Language ability and context of learning, *Language Testing*, 2013.
[23] J. E. Purpura, Assessing communicative language ability: Models and their components, in *Encyclopedia of Language and Education*, E. Shohamy and N. H. Hornberger (eds.), 2nd Edition, NY, Springer, 2010.
[24] ETS, *Internet-Based Test of English as a Foreign Language*, http://www.ets.org/toefl/ibt/about, 2012.
[25] F. M. Lord, The self-scoring flexilevel test, *Journal of Educational Measurement*, vol.8, pp.147-151, 1971.
[26] W. J. van der Linden and P. J. Pashley, Item selection and ability estimation in adaptive testing, in *Computerized Adaptive Testing: Theory and Practice*, W. J. van der Linden and C. W. Glas (eds.), Netherlands, Kluwer Academic Publishers, 2000.
[27] W. A. Sand, B. K. Water and J. R. McBride, *Computerized Adaptive Testing: From Inquiry to Operation*, American Psychological Association, Washington, DC, 1997.
[28] H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg and D. Thissen, *Computerized Adaptive Testing: A Primer*, Lawrence Erlbaum Associates Publish, Hillsdale, NJ, 1990.
[29] R. J. Adams, M. R. Wilson and W.-C. Wang, The multidimensional random coefficients multinomial logit model, *Applied Psychological Measurement*, vol.21, pp.1-23, 1997.
[30] J. Hattie, *Decision Criteria for Determining Unidimensional and Multidimensional Normal Ogive Models of Latent Trait Theory*, Armidale, Australia, 1981.
[31] R. L. Mckinley and M. D. Reckase, MAXLOG: A computer program for the estimation of the parameters of a multidimensional logistic model, *Behavior Research Methods & Instrumentation*, vol.15, pp.389-390, 1983.

[32] M. D. Reckase, The difficulty of test items that measure more than one ability, *Applied Psychological Measurement*, vol.9, pp.401-412, 1985.
[33] W. J. van der Linden and R. K. Hambleton, *Handbook of Modern Item Response Theory*, Springer-Verlag Press, New York, 1996.
[34] J. Mulder and W. J. van der Linden, Multidimensional adaptive testing with optimal design criterion for item selection, *Psychometrika*, vol.74, pp.273-296, 2009.
[35] D. O. Segall, Multidimensional adaptive testing, *Psychometrika*, vol.61, pp.331-345, 1996.
[36] W.-C. Wang and P. H. Chen, Implementation and measurement efficiency of multidimensional computerized adaptive testing, *Applied Psychological Measurement*, vol.28, pp.295-316, 2004.
[37] D. O. Segall and K. E. Moreno, Development of the computerized adaptive testing version of the armed services vocational aptitude battery, in *Innovations in Computerized Assessment*, F. Drasgow and J. B. Olson-Buchanan (eds.), Hillsdale, NJ, Lawrence Erlbaum Associates, 1999.
[38] J. de la Torre and H. Song, Simultaneous estimation of overall and domain abilities: A higher-order IRT model approach, *Applied Psychological Measurement*, vol.33, pp.620-639, 2009.
[39] M. D. Reckase, A linear logistic multidimensional model, in *Handbook of Modern Item Response Theory*, W. J. van der Linder and R. K. Hambleton (eds.), New York, Springer, 1996.
[40] J. de la Torre, H. Song and Y. Hong, A comparison of four methods of IRT subscoring, *Applied Psychological Measurement*, 2011.
[41] Y. Sheng and C. K. Wikle, Bayesian multidimensional IRT models with a hierarchical structure, *Educational and Psychological Measurement*, vol.68, 2008.
[42] H. Song, *A Higher-Order Item Response Model: Development and Application*, Ph.D. Thesis, Educational Statistics and Measurement, The State University of New Jersey, 2007.
[43] M. D. Reckase, *Multidimensional ITem Response Theory*, New York, Springer, 2009.
[44] H.-Y. Huang, P.-H. Chen and W.-C. Wang, Computerized adaptive testing using a class of high-order item response theory models, *Applied Psychological Measurement*, vol.36, pp.689-706, 2012.
[45] OECD, *PISA 2003 Technical Report*, OCED, Paris, 2005.
[46] Council of Europe, *The Common European Framework of Reference for Languages*, Cambridge University Press, Strasbourg, 2001.
[47] D. O. Segall, Principles of multidimensional adaptive testing, in *Computerized Adaptive Testing: Theory and Practice*, W. J. van der Linden and C. W. Glas (eds.), Netherlands, Kluwer Academic Publishers, 2000.
[48] F. M. Lord, Multidimensional computerized adaptive testing in a certification or licensure context, *Applied Psychological Measurement*, vol.20, pp.389-404, 1986.
[49] R. Mislevy, Bayes modal estimation in item response models, *Psychometrika*, vol.51, pp.177-195, 1986.