The Chinese as Second Language Multidimensional Computerized Adaptive Testing System Construction

Hsuan-Po Wang, Bor-Chen Kuo, Rih-Chang Chao, Ya-Hsun Tsai

***Abstract*—**With rising demand of Chinese as Second Language (CSL) learning, Chinese Proficiency Test became more and more popular recently. There are several major proficiency tests with paper-pencil (P&P) formats for Chinese learners including Taiwan's test of proficiency-Huayu (TOP-Huayu), the mainland's Hanyu Shuiping Kaoshi (HSK), and America's Scholastic Assessment Test (SAT). In this study, Common European Framework Reference (CEFR) is applied and CSL Proficiency Index is used as guidelines to develop a multidimensional computerized adaptive testing (MCAT) system for enhancing the CSL proficiency test. This research collected empirical data via the computerized based test (CBT) followed by developing and conducting a simulation study on a MCAT system. The proposed system provides a framework of using item response theory (IRT) as the ability scoring method and applies to the process as a MCAT. In addition, this research will also go through the evaluation of the effectiveness of the process on MCAT system. There were 658 empirical data collected from Grace Christian Collage in Philippine on September 2009. At the end of this research the result indicated that recommend CSL MCAT System applied MAP as the ability estimation method for this MCAT System. The interface of the MCAT system is also present at this research.

***Keywords*—**Chinese as second language Proficiency Test, the Common European Framework of Reference for Languages, Computer Based Test, Multidimensional Computerized Adaptive Test

I. Introduction

The trend of globalization of the multi-language skill learning has induced the impact on the increase of the numbers of countries to pay more attention to their education systems. One of the key affects was due to the economic boost in China which raises the demand to learn CSL. This demand was soon spread out all over the word including those governments in Americas, Europe, Asia, and other regions as well. There were a lot of countries established official organizations to promote CSL learning domestically. Indirectly this has increased the demand of CSL proficiency test to justify their proficiency. So there are many countries has launched their own CSL proficiency test and some of them even had conducted CSL computerized test for years such as SAT Chinese Listening subject and Advanced Placement (AP) Chinese Language and Culture subject both conducted by Collage Board in United States (College Board, 2010). Those different types of CSL proficiency test have created a discrepancy and incomparability on ability scale. The differences of ability scale have confused CSL learners not only in their ability justification but also in their curriculum engagement. Recently CEFR has been adopted by many different proficiency tests as the reference such as: 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 (TestDaF), Diplôme D'Etudes en Langue Française (DELF) etc. (Kecker & Eckes, 2007; Tannenbaum & Wylie, 2005). So this research will apply CEFR as a reference basis for conducting the CSL proficiency test.

The major CSL proficiency tests still adopted P&P test as the main way of testing. Because of advances in computer technology available, the CBT had been gradually become an important development project to replace the traditional P&P test in the near future such as TOP-Huayu, HSK, C. Test etc. However, it is not appropriated to conduct a computerized adapted test (CAT) with items including different domains because of the primary assumption of IRT unidimensionality. To address the problems, multidimensional IRT (MIRT) had been proposed and applied into the process as a MCAT (Hattie, 1981; Hsieh, Shih, & Chen, 2008; Luecht, 1996; Mckinley & Reckase, 1983; Reckase & Mckinley, 1991; Segall, 1996; Shih & Wang, 2007; Sympson, 1978). This research collected empirical data via CBT followed by developing and conducting a simulation study on a MCAT system. In addition to this, this research will also evaluate the effectiveness of the process on MCAT system.

All the CSL proficiency tests are still on a P&P testing style. There are no relative studies or papers available regarding of CSL MCAT

construction. The benefit and contribution of the research is to develop a MCAT system on a basis of CEFR for CSL proficiency test.

II. Application of MIRT and MCAT System Construction

This research will base on CEFR and develop a MCAT system for CSL proficiency test. There were 658 empirical data collected from Grace Christian Collage in Philippine on September 2009 followed by a simulation study via the MCAT process. This research will also compare the difference of the ability estimating methods between maximum likelihood estimation (MLE) and expected a posteriori (EAP) and maximum a posteriori (MAP) as a result to evaluate the effectiveness of the process on MCAT system. The procedures applied CEFR, MIRT and MCAT used in this article consist of the following steps.

A. *The Common European Framework of Reference for Languages, CEFR*

There are three major references used for languages learning, teaching and assessment: CEFR, American Council on the Teaching of Foreign Language (ACTFL) and Canadian Language Benchmarks (CLB). However, different references used by different CSL proficiency tests results a discrepancy and incomparable ability scale with each others. This ability scale different confused CSL learners not only in their ability justification but also in their curriculum engagement. The CEFR is intended to overcome the barriers to communication among professionals working in the field of modern languages arising from the different educational systems in Europe. It provides the means for educational administrators, course designers, and teachers, teacher trainers, examining bodies, etc., to reflect on their current practice, with a view to situating and coordinating their efforts and to ensuring that they meet the real needs of the learners for whom they are responsible (Council of Europe, 2001). There are 41 countries in Europe adopted CEFR as reference basis to design their foreign language curriculum, teaching methods and materials, teacher training and assessment tools. In addition, Hong Kong, Japan, New Zealand, Australia, Chile, Mexico, Colombia, Canada and other countries outside of Europe all use CEFR as reference for language test criteria.

CEFR categorized communication activities into four domains. There are productive activities and strategies, receptive activities and strategies, interactive activities and strategies, and mediating activities and strategies. Tsai (2009) based on CERF modified those four communication activities domains for CSL learners. The reason for modified those four domain contents was due to the language differences between English and Mandarin itself. CEFR constructed a detail, clear and coherent structure in able to providing a clear guidelines for different level CSL learners. Therefore, the research will apply CEFR as a reference basis for conducting the CSL proficiency test.

B. *Multidimensional Item Response Theory, MIRT*

The items in a multidimensional test can classify this test into two different types. One is called between-item multidimensional test and the other is called within-item multidimensional test. The difference between these two types of test is the domain(s) of one item that is measured during the test. Each Item in between-item multidimensional test measured one domain only. If one item in a test measured more than one domain at the same time, then this test was classified as a within-item multidimensional test (Adams, Wilson, & Wang, 1997). Most of the current used MIRT models are extend from unidimensional IRT models. They are multidimensional two parameters model (M2PL) proposed by Mckinley and Reckase (1983), multidimensional three parameters model (M3PL) proposed by Hattie (1981) and Sympson (1978), multidimensional version of the partial credit model (M-2PPC) proposed by Yao and Schwarz (2006) and MRCML model proposed by Adams, Wilson and Wang (1997).

Since the numbers of parameter and type of scoring method are different, MIRT model had extended several different types of models. MRCML model is the one had been widely used recently. The reason of its popularity because MRCML able to measure the partial credit data for a test with multidimensional domains and if

data contaminated the local independent assumption, MRCML can analyze data via testlet format.

MRCML model is a MIRT model extended from Rasch model (Hoskens & De Boeck, 1997; Wang, Wilson, & Cheng, 2000; Wilson & Adams, 1995). The definition of MRCML model was described bellowed equation 1:

$$P(X_{ik}=1;A,B,\xi \mid \theta)=\frac{\exp(b'_{ik}\theta+a'_{ik}\xi)}{\sum_{k=1}^{K_i}\exp(b'_{ik}\theta+a'_{ik}\xi)} \quad (2)$$

where X_{ik} represent the participant's response; K_i represent the score category of i^{th} item; θ represent the parameter matrix of participant's multidimensional ability; ξ represent the vector of item parameters; b_{ik} represent the scoring matrix of k^{th} response in i^{th} item; a_{ik} represent the designing matrix of k^{th} response in i^{th} item.

C. *Multidimensional Computerized Adapted Test, MCAT*

CAT system immediately estimated participant's ability via his or her response on the items, and selected the next item which closer to the participant's ability for them to answer during the test processing. This procedure had been proved that CAT system can use fewer items to estimate the participant's ability compared with the traditional paper-pencil test did with same accuracy.

There are two major CAT systems. One is called Unidimensioanl CAT (UCAT) and the other is called MCAT. Since the CSL proficiency test contains different domains, the objective of this research is to develop a MCAT system. Based on IRT, CAT system can calibrate and estimate a participant's ability as soon as the response of the item had collected. The CAT system will then select the next item which closer participant's ability for him or her to answer during the test conducting. According to this processing, although each participant answered different items for a same test, they took items that were much closer to his or her ability in order to achieve more accuracy in their ability estimation. The MCAT system had been developed in this search was based on IRT. This system contains five basic components: Item Bank Development, Test Starting Point evaluation, Parameters Estimating Process, Item Selection Strategy, and Test Termination Criteria. Figure 1 indicated the MCAT procedures used in this research.

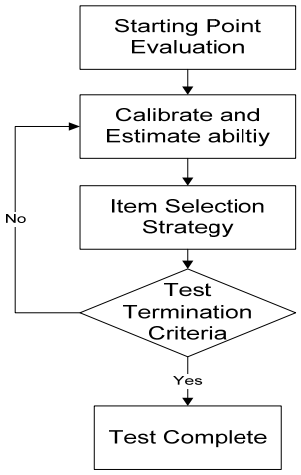

Figure 1. MCAT procedures used in this research

III. Construct a MCAT System

A. *Implementation Process*

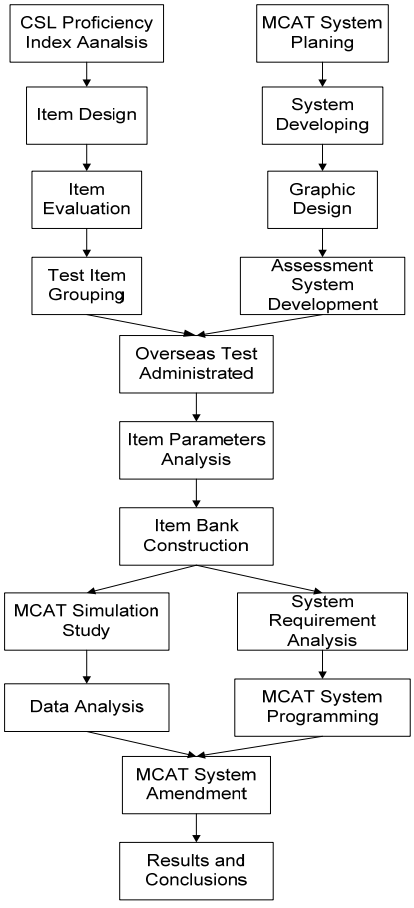

Figure 2. MCAT System Implementation Procedure

The implementation procedures of MCAT system for CSL proficiency test had indicated below shown in Figure 2.

Firstly, processing on system analysis and design according to this research objective. Secondly, conduct MCAT system programming based on the user module. Thirdly, calibrate and analyze item parameters and feed back to MCAT system together with participants' responses. Last but not at least, amends and implements the MCAT system according to the result of this testing process.

B. *MCAT System Analysis and Design*

The framework of MCAT system in this research had constructed into two sides as shown in Figure 3. One is the client side and the other is the server side. The client side is the interface of user operation. Users can login into system via browse through HTML website. The Linux CentOS 5 operation system and MYSQL database in the server side were used for data processing and stored participant's information, item response, item bank etc. The software used for the web server is Apache. The functions of each module and their link procedures to the database were programmed by PHP. This PHP program was also implemented with other software such as HTML and JavaScript to fulfill the functions needed for each module.

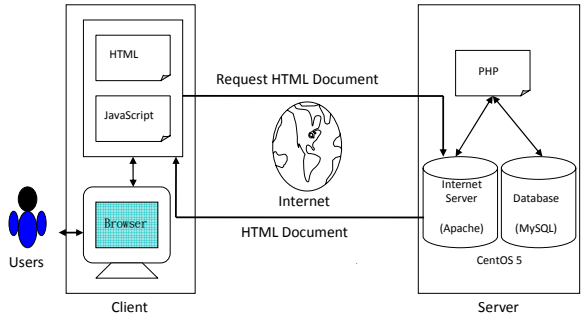

Figure 3. MCAT System Framework

C. *MCAT Processing Flow*

The MCAT system designed in this research will assign an item which corresponding to an initial ability prior defined for each participant. As soon as the participant completed this item, the MCAT system will utilize MAP, MLE and EAP three different methods for calibrating and estimating participant's ability respectively. The item selection strategy is based on the Maximum Information Method. All the items in the item bank will control by the testing item exposure rate. That information will be provided for the guideline in MCAT system to indicate which items should or should not be chosen during a test. The test will reach at the end if item selection strategy met the test termination criteria. The MCAT system processing flow display in Figure 4 shown below.

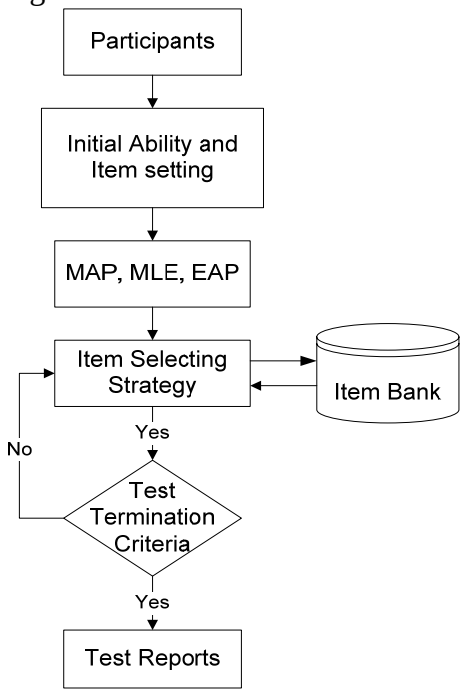

Figure 4. MCAT Processing Flow

D. *System Module Design*

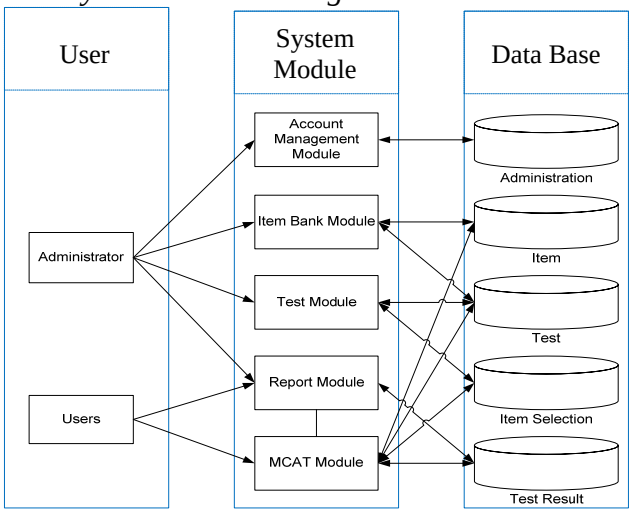

Figure 5. MIRT Based MCAT modular Structure

This MCAT system not only provide testing module but also supply other functional module such as MCAT Module, Account Management Module, Item Bank Module, Report Module etc.

For example, the examinees are able to participate in the test through the MCAT System via MCAT Module, and check on the response on each item and the test report via Report Module as well. Administrator can manage the examinees' accounts by adding a new account, amend passwords or even deleted an old account via Account Management Module. He or she can also maintain the item bank by adding or amend the items via Item Bank Module. The Test Module helps administrator assign different test booklets to different examinees. The MIRT based MCAT system was indicated in Figure 5.

a. Account Management Module

Participants can modify and inquire their personal information only via Account Management Module. The Administrator is allowed to add or modify including import or export some or all participants' account information by queried the database via a specific command.

b. Item Bank Module

The Item Bank manager can add or edit test structure, select test level or test type, and assign booklet to participant as well via Item Bank Module; He or she can also add or edit the test instruction, items and block and mark them for modification impossible. The item parameters (e.g. item difficulty parameter) import and export functions had implement into this Module.

c. Test Module

Teat Module monitor the loading and the booklets assignment actual used of the site and assign to every participants in each class for participating the test. Therefore, the Administrator can assign or cancel any test booklet with a specific test type to any classes. He or she can also select the item response model, item selection strategy, ability estimation methods, and the maximum lest length as well via the functions in the Test Module.

d. Report Module

Participants and the Administrator both can confirm their personal information, check on their test results and query the diagnostic reports via Report Module.

e. MCAT Module

The system will utilize the item response model and the ability estimation method which assigned in the beginning of the test to calibrate and estimate the each participant's temporarily ability. The next item will be selected by the item selection strategy based on this participant's temporarily ability. This procedure will go all the way through and stop as soon as it met the Test Termination Criteria. The MCAT Module applied UIRT model as its item response model and applied Maximum Information method on item selection strategy. In addition, the MCAT Module applied MLE, MAP and EAP for the participant's ability calibration and estimation.

E. *Experimental Design*

a. Sample collecting

This research is based on CEFR as the reference to design A2 level CSL Proficiency Test. This test content included three different domains. There are productive activities and strategies, receptive activities and strategies, and interactive activities and strategies. There were total of 59 items in this test. This research collected empirical data via CBT which designed and developed by Kuo and Zheng (2010) in National Taichung University. There were 658 empirical data collected from 7th to 10th grade students of Grace Christian Collage Philippine on September 2009.

b. Ability Calibration and Estimation

The more items in the simulation study of MCAT system, the lesser root mean squared error (RMSE) on all MLE, MAP or EAP methods of ability calibration and estimation.

c. The Effectiveness Index of Ability Estimation

The ability estimation accuracy and the effectiveness of MCAT system were evaluated by the difference between real ability and estimated ability, RMSE. RMSE is defined as follows equation 2:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_i - \theta_i)^2} \quad (2)$$

where θ_i represent the i^{th} participant's ability estimation at the end of the test. MCAT System defined this value as the i^{th} participant's real ability; $\hat{\theta}_i$ represent the ith participant's temporarily ability estimation after partial items had been responded; N represent total number of participants.

IV. Results and Conclusions

A. *MCAT System Interfaces*

a. Administration Interfaces

The administration interface is a very user friendly implementation process. It can be imported file one time in excel format or be assigned for more than one participants to participate the test or to modify the participants' information. Each participant had to log in the system followed by fill up all the questions on the questionnaire and selected the grade and type of the test he or she want to participant with. The Administration Interface of MCAT system was shown as below as figure 6.

Figure 6. Administration Interfaces of MCAT System

b. Report Interfaces

The report interface of MCAT System is able to check on the test result on each participant's or class including administer time, response sequence, each class's statistics information. In addition to this, this interface can also report the average correct response rate on the basis for each student or class. The Report Interface of MCAT system is display below figure 7.

Figure 7. Report Interfaces of MCAT System

B. *Simulation Study on MCAT System*

A complete set of item response empirical data from CSL proficiency test was collected via CBT and followed by a simulation study via MCAT System. The objective of these simulation study procedures was to find out the variation of RMSE while applying different ability estimation methods among different ability domains under the test with different number of items.

Apply MAP as a method on ability calibrating and estimation in the simulation study indicated that, according to the results plot in figure 8, as soon as the participant completely responded 10 items in CSL proficiency test, the RMSE of receptive activities and strategies domain was closed to 0.3, and the RMSE of productive activities and strategies, and receptive activities and strategies domains were both closed to 0.45. When the participant completed responded 20 items, the RMSE in all three domains were less than 0.3.

Applied MLE as a method on ability calibrating and estimation in the simulation study indicated that, according to the results plot in figure 9, as soon as the participant completely responded 45 items in CSL proficiency test, the RMSE of receptive activities and strategies domain was less than 0.3 as MAP method did, and the RMSE of productive activities and strategies, and receptive activities and strategies domains were closed to 0.6 and 0.5 respectively which was much worse compared with MAP method did.

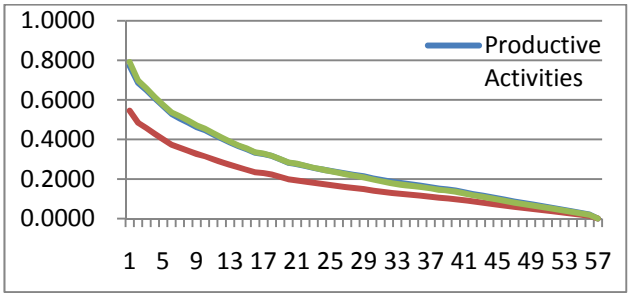

Figure 8. The variation of MAP estimation among three domains

While applied EAP as a method on ability calibrating and estimation in the simulation study indicated that, according to the results plot in figure 10, as soon as the participant completely responded 30 items in CSL

proficiency test, the RMSE of receptive activities and strategies domain was less than 0.3 as MAP and MLE method did, and the RMSE of productive activities and strategies, and interactive activities and strategies domains were closed to 0.4 and 0.45 respectively which was very close to the result from MAP method. When the participant completely responded up to 40 items, only the RMSE of interactive activities and strategies domain were a little more than 0.3. The RMSE in other two domains were both less than 0.3.

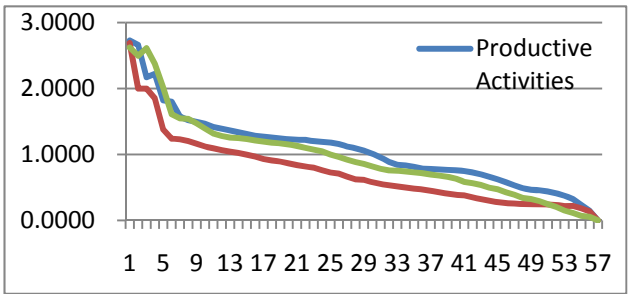

Figure 9. The variation of MLE estimation among three domains

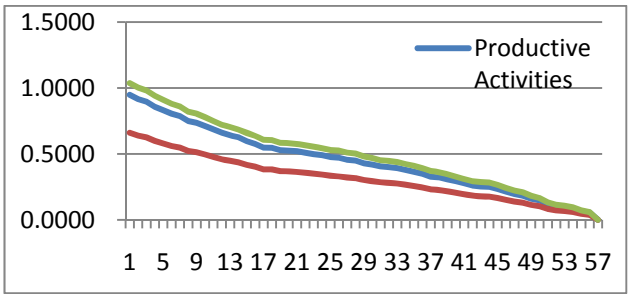

Figure 10. The variation of EAP estimation among three domains

According to Chen (2006) MAP and EAP method had higher reliability and less measurement error compared with MLE did in the MCAT system. In addition to this, the more domains contained in a test, the longer administration time EAP method was needed on ability calibration and estimation. Therefore, this research recommends applying EAP method as the ability estimation method in MCAT System when there is lesser domain contained in the test. Alternatively this research recommends applying MAP method as the ability estimation method in MCAT System when there are more domains contained in the test. Overall speaking MAP method performed better than EAP and MLE method did on the ability estimation in MCAT System. Therefore, this research will apply MAP as the major method on the participants' ability calibration and estimation in MCAT System for the experimental study in the near future.

C. *Future direction of study and suggestion*

The MCAT System constructed for CSL proficiency test in the research allows the participant to check on the results of the test after login into the Administration Interfaces. As soon as the connection established between system and data base was completed, participants can quarry the test results and receive the feedback immediately from the MCAT System. MCAT System also developed several modules available for administrator, such as Item Bank module and Test Module etc. For example, the administrator can add or amend item(s) in his or her convenient time within or outside the item bank. Addition to this, the administrator can assign test type(s) and/or booklet(s) via the functions in Test Module. For example, In Test Module the administrator can assign different test type(s) and booklet(s) for two different group participants at different time to participate the test.

All the CSL proficiency tests are still on a P&P testing style. There is no CSL proficiency test constructed in any MCAT system so far. The benefit and contribution of the research was not only the development of a MCAT system but also constructed a CSL Proficiency Test on a basis of CEFR. However, there are three directions list below vital for the future study.

1) Modify MCAT System: enable to fit different item style such as non-multiple choice item for CSL proficiency test.
2) Simple Chinese version: enable more users in different countries to participate the CSL proficiency test via MCAT System.
3) Function Enhancement: enable more functional selections on initial ability setting criteria and item selection criteria. Apply item exposure rate control criteria for MCAT System as well.

REFERENCES

[1] College Board (2010). Chinese with Listening. Retrieved May 20, 2010, from http://www.collegeboard.com/student/testing/sat/lc_two/chinese/chinese.html?chinese

[2] Kecker, G., & Eckes, T. (2007). *Linking the TestDaF to the CEFR: The case of writing proficiency.* Paper presented at the Fourth

Annual Conference of EALTA. Retrieved August 4, 2009, from http://www.ealta.eu.org/conference/2007/docs/pres_sunday/Kecker&Eckes.pdf

[3] Tannenbaum, R. J., & Wylie, E. C. (2005). *Mapping English proficiency test scores onto the Common European Framework* (TOFEL Research Rep. NO. RR-80). Retrieved August 4, 2009, from http://www.ets.org/Media/Research/pdf/RR-05-18.pdf

[4] Hattie, J. (1981). *Decision criteria for determining unidimensional and multidimensional normal ogive models of latent trait theory.* Armidale, New South Wales, Australia: The University of New England, Center for Behavioral Studies.

[5] Hsieh, C. L., Shih, C. L., & Chen, C. T. (2008). *Efficiency and responsiveness to change of multidimensional computerized adaptive testing of movement in stroke patients*. Parper presented at the Pacific Rim Objective Measurement Symposium, Japan, Tokyo.

[6] Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement*, *20*, 389-404.

[7] Mckinley, R. L., & Reckase, M. D. (1983). MAXLOG: A computer program for the estimation of the parameters of a multidimensional logistic model. *Behavior Research Methods & Instrumentation*, *15*, 389-390.

[8] Reckase, M. D., & Mckinley, R. L. (1991). The discrimination power of items that measure more than one ability. *Applied Psychological Measurement*, *15*, 361-373.

[9] Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, *61*, 331-345.

[10] Shih, C. L. & Wang W. C. (2007). *Comparison of Item Selection Strategies in Multidimensional Computerized Adaptive Testing*. Pacific Rim Objective Measurement Symposium, Taiwan.

[11] Sympson, J. B. (1978). A model for testing with the multidimensional items. In D. J. Weiss (Ed.). *Item response theory and computerized adaptive testing conference proceedings.* MN: University of Minnesota press.

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

[13] Adams, R. J., Wilson, M. R. & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*, 1-23.

[14] Tsai, Y. H. (2009). *CSL Teaching Materials Construction*. Taipei: Zhong Zheng, 2009.

[15] Yao, L., & Schwarz R. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed format tests. *Applied Psychological Measurement*, *30*, 469-492.

[16] Hoskens, M., & De Boeck, P. (1997). A parametric model for local dependencies among test items. *Psychological Methods*, *2*, 261-277.

[17] Wang, W., Wilson, M., & Cheng, Y. (2000). *Local Dependence between Latent Traits when Common Stimuli are Used.* Paper presented at the International Objective Measurement Workshop, New Orleans, LA.

[18] Wilson, M., & Adams, R. J. (1995). Rasch models for item bundles. *Psychometrika*, *60*, 181-198.

[19] 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.

Author

Hsuan-Po Wang was born in Nantou, Taiwan. His day of birth is on July 1st, 1982. He earned his master degree from Graduate Institute of Educational Measurement and Statistics (GIEMS) at National Taichung University, Taichung, Taiwan.. He is currently a PH. D candidate of GIEMS at National Taichung University, Taichung, Taiwan. His research interests are specialized in the areas of proficiency measurement and assessment for learning Chinese as Second Language (CSL).

Co-authors

Bor-Chen Kuo received the Ph.D. degree from the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, in 2001. He is currently a Professor with the Graduate Institute of Educational Measurement and Statistics (GIEMS), National Taichung University, Taichung. His research interests are pattern recognition, remote sensing, image processing, and nonparametric functional estimation.

Rih-Chang Chao was born in Kaohsiung, Taiwan. His day of birth is on April 7th, 1963. He earned his master degree from Department of Applied Probability and Statistics at Northern Illinois University in 1992. He is currently a PH. D candidate of the Graduate Institute of Educational Measurement and Statistics (GIEMS) at National Taichung University, Taichung, Taiwan. He was working for AC Neilson Consultant Company at Taiwan subsidiary as a senior manager in Department of Client Solutions. He possessed local and foreign profession education background, and have working experience in USA and Taiwan in the field of Market Research. With more than 15 years of working experience he had gained a boarder view in international business development, theoretical and practical analyze capability, most importantly he is very familiar in the Taiwanese cultural, business and commercial interaction, especially in handle and manage the complexity between vendor and customer relationship. In 2008, he has started teaching in National Taiwan Normal University, Taipei, Taiwan and become a lecturer of Mathematics and Science department. Major focus on his research is on the language testing. His research interests are specialized in the areas of proficiency measurement and assessment for learning Chinese as Second Language (CSL).

Ya-Hsun Tsai is a Professor and Department Head of Applied Chinese Language and Literature at National Taiwan Normal University, Taiwan. She has been in this position since 2008. Before joined NTNU, she was the Department Head of Applied Linguistics and Language Studies at Chung Yuan Christian University. Professor Tsai has published many articles, journal and books in the areas of Teaching Chinese as Second Language (CSL). She is the author of 『Teaching Material Construction for Learning Chinese as Second Language』 and co-authored a book on the 『Introduction of Teaching Chinese as Second Language』. She also had been officially nominated as one of the committees by both Ministry of Education and Overseas Compatriot Affairs Commission and participated with several CSL related projects. Her current research interests are also including CSL Testing and Assessment, CSL Teaching Syllabus Design, and Standard Setting for CSL vocabulary.