A HO-IRT BASED DIAGNOSTIC ASSESSMENT SYSTEM WITH CONSTRUCTED RESPONSE ITEMS

Chih-Wei Yang
Graduate Institute of Educational Measurement and Statistic,
National Taichung University of Education, Taiwan
Chihwei_yang@hotmail.com

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

Chen-Huei Liao
Department of Special Education, National Taichung University of Education, Taiwan
chenhueiliao@gmail.com

ABSTRACT

The aim of the present study was to develop an on-line assessment system with constructed response items in the context of elementary mathematics curriculum. The system recorded the problem solving process of constructed response items and transfered the process to response codes for further analyses. An inference mechanism based on artificial intelligence was implemented with the system to diagnose the bugs in the problem solving process automatically. To examine the performance of the system, a "Multiplication of Fraction" test was constructed and administered to 158 six graders in Taiwan. The results showed that the mean of classification accuracies of the bugs is above 97%, which implies that the proposed system identifies leaning bugs accurately and efficiently. In addition to bug identification, a high-order item response theory (HO-IRT) was applied to estimate the overall and domain abilities. The correlations between the abilities estimated with HO-IRT and the number of bugs were highly correlated, which suggests that the more learning bugs children possessed the lower his/her mathematic abilities would be.

Keywords: constructed response item, computerized test, automated scoring, high-order item response theory

INTRODUCTION

Constructed response (CR) items are open ended, short answer questions that elicit students' higher-level cognitive abilities and are beneficial to evaluate complex concepts or skills such as problem solving (Martinez & Bennett, 1992; Zenisky & Sireci, 2002; Bacon, 2003; Williamson, Bejar, & Sax, 2004, Kuechler & Simkin, 2010). The solving process of CR items involves multiple steps, which explicitly demonstrates how the final answers are derived, and sometimes, students are required to provide explanation in writing to support his / her answers. The responses of CR items will later be classified into different response types, and the scores will be given according to the actual performance. Also, due to different steps are involved in every single CR item, a unique rubric for scoring is required for each item.

The constructed response items have been used by some large-scale assessments, such as NAEP, PISA and TIMSS (National Assessment Governing Board, 2005; Olson, Martin, & Mullis, 2008; OECD, 2005; Parshall, Davey, & Pashley, 2002). The NAEP example items can be found on the website. (http://nces.ed.gov/nationsreportcard/itmrlsx/search.aspx?subject=mathematics). Taking NAEP scorning method as an example, the answers were rated into five response levels based on the completeness of the answers, which are extended, satisfactory, partial, minimal, and incorrect, and the criteria for each level are clearly described. Traditionally, human scorers of CR items must be well-trained to strictly follow the criteria to make sure that the responses are scored consistently. However, the training process of human scorers is time-consuming and economically inefficient. Although constructed response items provide sufficient and crucial information of student's learning process, the high cost of time and money involves in manual grading remained challenging to educators and researchers (Attali, & Burstein, 2006; Attali, Powers, Freedman, Harrison, & Obetz, 2008).

In Figure 1, two examples of one of the available computerized tests with automated scoring system are illustrated. In these two items, the input formats and variations of items are limited, and the system does not record the problem solving process. Therefore in the present study, an on-line assessment system was developed to compensate the current practice of CR items. The system allows the solving process of CR items to be recorded completely and learning bugs (error patterns) to be analyzed automatically and instantly. Moreover, the overall and domain abilities are estimated simultaneous by higher-order item response theory.

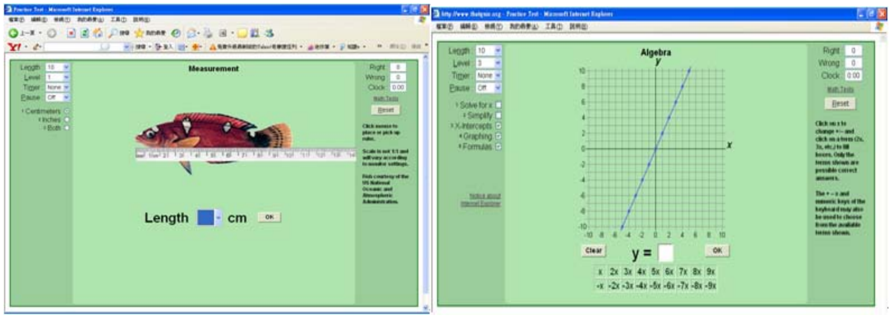
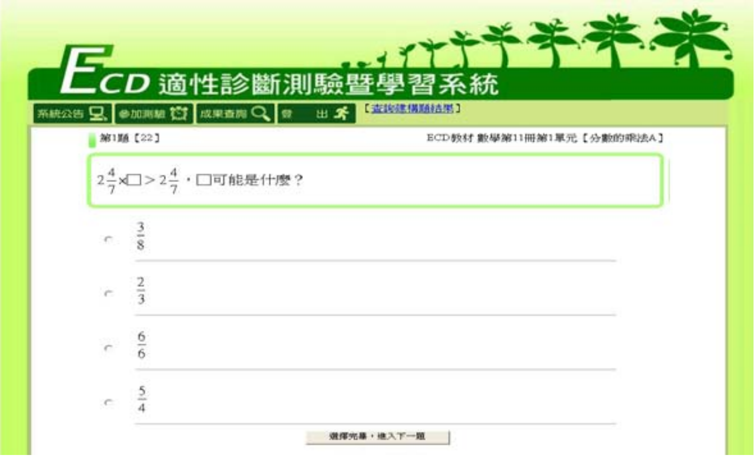


Figure 1 Computerized constructed response items
(Thatquiz, http://www.thatquiz.org/)

METHOD

Computerized Test Development

In figure 2 and figure 3 the test interfaces of multiple choice and constructed response items are shown. The responses, like the equations or the fractions, can be inputted by using the tool the system provided adaptively for each CR item. The inputted equations will be displayed in the response area and recorded by LaTeX format. Table 1 is an example of an inputted equation and the corresponding codes in the database.



第1題【22】 ECD教材 數學第11冊第1單元【分數的乘法A】

$2\frac{4}{7}\times\square > 2\frac{4}{7}$，□可能是什麼？

○ $\frac{3}{8}$

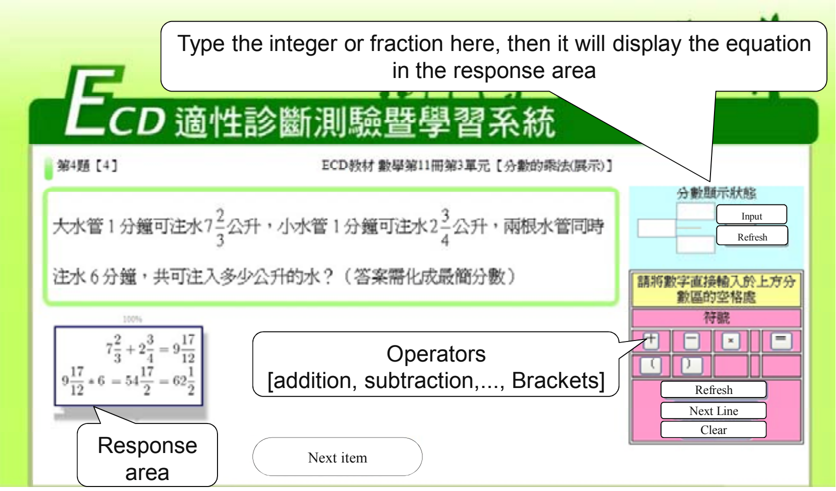
○ $\frac{2}{3}$

○ $\frac{6}{6}$

○ $\frac{5}{4}$

選擇完畢，進入下一題

Figure 2 Interface of multiple choice items


ECD 適性診斷測驗暨學習系統
分數顯示狀態
Input
Refresh
請將數字直接輸入於上方分數區的空格處
符號
Refresh
Next Line
Clear
Next item

Type the integer or fraction here, then it will display the equation in the response area

第4題【4】 ECD教材 數學第11冊第3單元【分數的乘法(展示)】

大水管 1 分鐘可注水$7\frac{2}{3}$公升，小水管 1 分鐘可注水$2\frac{3}{4}$公升，兩根水管同時注水 6 分鐘，共可注入多少公升的水？（答案需化成最簡分數）

$$7\frac{2}{3}+2\frac{3}{4}=9\frac{17}{12}$$
$$9\frac{17}{12}*6=54\frac{17}{2}=62\frac{1}{2}$$

Operators
[addition, subtraction,..., Brackets]

Response area

Figure 3 Interface of constructed response items

Table1 an example of the response format

Response pattern	$12\frac{8}{15}\times 1\frac{1}{4}=\frac{180}{15}\times\frac{5}{4}=\frac{45}{3}=15$
Database	12\\frac{8}{15}* 1\\frac{1}{4} = \\frac{180}{15}* \\frac{5}{4} = \\frac{45}{3} = 15\\frac{}{}

The analysis of constructed response item

One of the purposes of this study was to develop an automated analysis process for CR items to diagnose error patterns. There were two parts in the analysis process; first, some rules were used to build a decision tree and to classify the responses into several categories. Second, the prototypes of error patterns were compared to responses of the participants by using the block-based matching analysis.

For example, in item 27, three rules are involved in the decision tree (see Figure 4)

Rule 1: check the status of the response area. If the response area is blank, then code 99 will be given; otherwise, rule 2 will be applied.

Rule 2: examine the correctness of the first equation in the participant's response by comparing with the correct equations preset in the system. If the first equation is correct, then apply to Rule 3a; otherwise apply to Rule 3b

Rule 3a: check the correctness of the final answer. If the student's final answer is correct, then the system will record that the student has answered this item correctly; otherwise the block-based matching analysis is applied to find the best fit error pattern from the prototypes of Bug1 to Bug9.

Rule 3b: if the error occurs due to fraction addition instead of fraction multiplication, then the block-based matching analysis will find the best fit error pattern from the prototypes of Bug10 and Bug11. Otherwise, the error pattern will refer to Bug 12, which implies that the student did not understand the problem.

The first step of block-based matching analysis is to decompose the student's response into blocks without operators, then compare these blocks with bugs' prototypes and find the best fit error patterns.

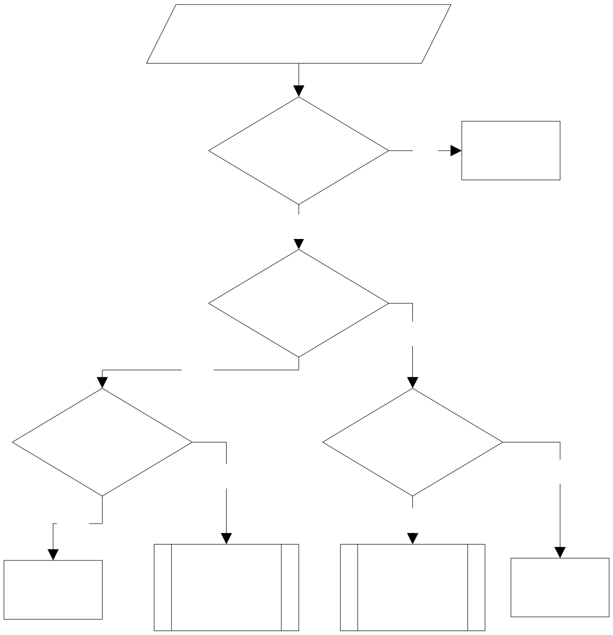


Figure 4 Analysis flow chart of item 27

Higher-Order Item Response Theory

A hierarchical structure of the higher-order item response theory (HO-IRT) framework involves the overall ability at the first layer and multiple domain abilities at the second layer. This framework has been well adopted at large-scale assessment settings. de la Torre and Song(2009) proposed a overall and multiple domain abilities simultaneously. de la Torre & Song(2009) and de la Torre & Hong(2010) show that parameter estimation by applying HO-IRT is more accurate and reliable than that by using traditional unidimensional IRT and multidimensional IRT separately. In this study, each domain of HO-IRT is considered to be unidimensional, and a domain-specific ability $\theta_i^{(D)}$ accounts for the performance of examinee i on domain d, where $d = 1, 2, ..., D$. The correlations between the different domain abilities are accounted for by positing a higher-order overall ability θ_i. Specifically, the domain abilities are linked to the overall ability via the linear function $\theta_i^{(d)} = \lambda^{(d)}\theta_i + \varepsilon_{id}$, where the $\lambda^{(d)}$ is the latent regression coefficient of the domain ability d on the overall ability, and ε_{id} is the error term which follows a standard normal distribution. Markov chain Monte Carlo (MCMC) method in the WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000) was used to estimate item and ability parameters simultaneously. In this framework, the response of examinee i to the jth item of the dth domain, $X_{ij}^{(d)}$, is a function of $\theta_i^{(d)}$ and the specific item characteristic via some IRT model. In this study, the multidimensional random coefficients multinomial logit (MRCML) (Adams, Wilson & Wang, 1997) model is employed, where the $\mathbf{b}_{ik}$, $\mathbf{a}_{ik}$ and $\boldsymbol{\xi}$ are the scoring matrix, design matrix and item parameter vector respectively. The probability of a response in category k of item j can be expressed as the following formulation.

$$\mathrm{P}(X_{jk} = 1; \mathbf{A}, \mathbf{B}, \boldsymbol{\xi} \mid \boldsymbol{\theta}) = \frac{\exp(\mathbf{b}_{jk}\boldsymbol{\theta} + \mathbf{a}'_{jk}\boldsymbol{\xi})}{\sum_{k=1}^{K_j} \exp(\mathbf{b}_{jk}\boldsymbol{\theta} + \mathbf{a}'_{jk}\boldsymbol{\xi})}$$

Test Description

A test of “Multiplication of Fraction” was developed based on the mathematics curriculum in Taiwan. One hundred and fifty-eight six graders were recruited from 4 elementary schools in Taiwan. There were 30 items in the test, including 26 multiple choice (MC) items and 4 constructed response (CR) items. In the present study (as shown in figure 5), the mathematical ability (overall ability) was assessed by the three domain abilities, in which conceptual knowledge was measured by MC item 1-4, procedural knowledge was measured by MC item 5-14, and problem solving was measured by MC 15-26 & CR 27-30.the test measured the mathematical ability and three domain abilities, conceptual knowledge, procedural knowledge and problem solving.

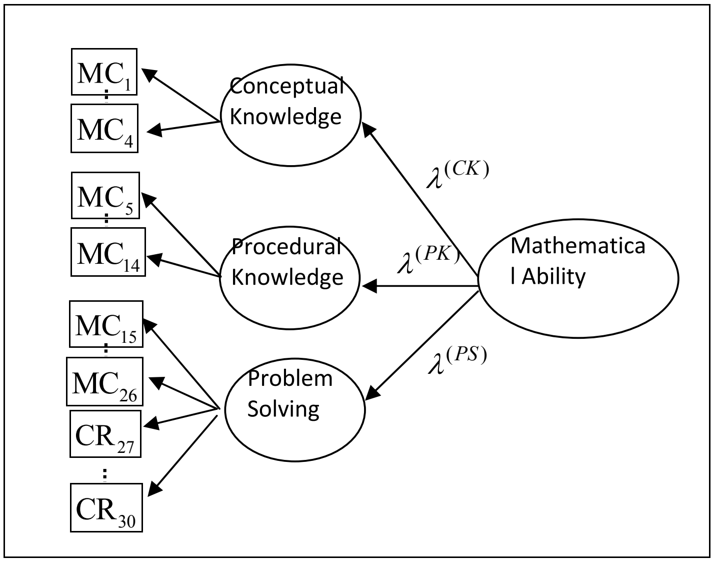


Figure 5 assessment framework of HO-IRT

RESULTS

The effectiveness of the error pattern diagnosis was examined by comparing the automated scoring results of the system to the human scoring results. In Table 2, the classification accuracy represents the percentage of the same diagnostic results by the system and human raters. The classification accuracies of constructed response items ranged from 94.94% to 99.37%, which means that the capability of the proposed system to diagnose student’s learning bugs is close to that of human raters.

Table 2 Classification accuracies of the proposed method for learning bugs

CR items	Classification accuracies	# of learning bugs in this item
27	99.37%	12
28	98.73%	11
29	94.94%	10
30	95.57%	9

Table 3 shows the estimated regression coefficients by using the MCMC method under the HO-IRT model. This result showed that the three domain abilities and the overall mathematics abilities were highly correlated.

Table 3 The regression coefficient of the domain abilities

variable	$\lambda^{(CK)}$	$\lambda^{(PK)}$	$\lambda^{(PS)}$
Mean of posterior	0.947	0.948	0.979

Table 4 shows the correlations between students' abilities and the number of learning bugs. The number of error patters were highly correlated with the overall mathematics ability (r = -0.901), whereas the correlations between the number of error patterns ranged from r =-0.896 to -0.907. The results showed that the more error patterns the student possessed, the lower his/her mathematic ability was. Therefore, the results provide evidence that the proposed system successfully and effectively identify students' error patterns.

Table 4 Correlations between the abilities and the number of bugs

	Domain			
Variable	MA	CK	PK	PS
The number of Bugs	-0.901	-0.896	-0.907	-0.898

Note: MA= Mathematical Ability; CK= Conceptual Knowledge; PK=Procedural Knowledge; PS=Problem Solving.

CONCLUSION

The present study developed an on-line assessment system with constructed response items. The results showed that the current system effectively and efficiently identifies student's learning bugs. Moreover, the utility of the system in the real class settings was observed.

The diagnostic assessment system with constructed response item provides precise diagnostic information, which not only provides the process of problem solving, but also identifies the difficulties children encounter in learning mathematics. The overall ability estimated by HO-IRT provides crucial information for important decisions such as rank-ordering the students and the domain abilities, which were also estimated by HO-IRT, identify students' strengths and weaknesses in the process of learning. In a computer-based testing environment, speed and efficiency are gained through automated scoring (Zenisky & Sireci, 2002), thus, this on-line assessment system not only generates diagnostic feedbacks instantly that may potentially aid teachers to direct students to more remedial instruction, it will also promote the behavior of self-study among children.

The present study provides evidence that the assessment system is helpful for both teachers and students. However, for solving the constructed response items, it is time consuming for children to input the equations on the computer screen with the system. Therefore, even though instructions were clearly explained and practice items were provided before testing, children who were lack of computer skills or those who were unfamiliar with the system, experienced difficulties key in required information. For future studies, different types of constructed response item that provide easier access of tools or gadgets, such as geometric items using a variety of onscreen drawing tools (Bejar, 1991;Williamson, Bejar, & Hone, 1999;Williamson, Hone, Miller,&Bejar, 1998), should be developed and investigated.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the support of this study by the National Science Council of Taiwan, under the Grant No NSC 97-2511-S-142-004 and NSC 99-2410-H-142-008-MY3.

REFERENCE

Adams, R., Wilson, M. and Wang, W. (1997) The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21,* 1-23.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater V.2. *Journal of Technology, Learning, and Assessment, 4*(3). Retrieved March 21, 2008, from http://escholarship.bc.edu/jtla/vol4/3/

Attali, Y., Powers, D., Freedman, M., Harrison, M., & Obetz, S. (2008). *Automated Scoring of Short-Answer Open-Ended GRE Subject Test Items.* (GRE Board Research Rep. No GRE-04-02).Princeton, NJ: ETS.

Bacon, D. R. (2003). Assessing learning outcomes: A comparison of multiple-choice and short answer questions in a marketing context. *Journal of Marketing Education, 25*, 31–36

Bejar, I. (1991). A methodology for scoring open-ended architectural design problems. *Journal of Applied Psychology, 76*, 522–532.

de la Torre, J., & Song, H. (2009). Simultaneous Estimation of Overall and Domain Abilities: A Higher-Order IRT Model Approach. *Applied Psychological Measurement, 33(8),* 620-639.

de la Torre, J., & Hong, Y. (2010). Parameter Estimation With Small Sample Size A Higher-Order IRT Model Approach. *Applied Psychological Measurement, 34(4),* 267-285.

Kuechler, W. L. and Simkin, M. G. (2010), Why Is Performance on Multiple-Choice Tests and Constructed-Response Tests Not More Closely Related? Theory and an Empirical Test. *Decision Sciences Journal of Innovative Education, 8*, 55-73.

Lunn, D.J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS -- a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing, 10*, 325--337.

Martinez, M. E., & Bennett, R. E. (1992). A review of automatically scorable constructed-response item types for large-scale assessment. *Applied Measurement in Education, 5*(2), 151-169.

National Assessment Governing Board (2005). 2005 National Assessment of Educational Progress Mathematics Assessment and Item Specifications. Washington, DC: Author.

Olson, J.F., Martin, M.O., & Mullis, I.V.S. (Eds.). (2008). TIMSS 2007 Technical Report. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

OECD (2005). PISA 2003 Technical Report. OCED. Paris.

Parshall, C. G., Davey, T., & Pashley, P. J. (2002). Innovative Item Types for Computerized Testing Computerized Adaptive Testing: Theory and Practice. In W. J. Linden & G. A. W. Glas (Eds.), (pp. 129-148): Springer Netherlands.

Williamson, D. M., Bejar, I. I., & Hone, A. S. (1999). 'Mental model' comparisons of automated and human scoring. *Journal of Educational Measurement, 36,* 158–184.

Williamson, D. M., Bejar, I. I., & Sax, A. (2004). Automated Tools for Subject Matter Expert Evaluation of Automated Scoring. *Applied Measurement in Education, 17*(4), 323-357.

Williamson, D. M., Hone, A. S., Miller, S., & Bejar, I. I. (1998, April). *Classification trees for quality control processes in automated constructed response scoring*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Zenisky, A.L., & Sireci S.G. (2002). Technological Innovations in Large-Scale Assessment, *Applied Measurement in Education, 15(4),*337-362