

Improved Effort-Moderated Item Response Model for E-learning Performance and Assessment

Cheng-Yin Tang and Tun-Wen Pai*

*Dept. of Computer Science and Engineering,
National Taiwan Ocean University, Taiwan, Republic of China*

**twp@mail.ntou.edu.tw*

Abstract: Item response model in psychological measurement is the most important evaluation method in contemporary education examination and its evaluation results can manifest in variable levels according to candidates' effort. This study mainly proposed a so-called Improved Effort-moderated (IEM) item response model, which primarily combined item response time in Effort-moderated (EM) item response model and precise self-defined answer indices to estimate proficiency level of the candidate. When there were 61,000 candidates within rapid answering behavior and random guess behavior, the capability estimated by IEM model showed smaller RMSE than by EM model and Three-parameter logistic model. The proposed model possesses more accurate capability evaluation than existing methods. IEM model inherited characteristics of EM model that originally outpaced Three-parameter logistic model, and indeed made up EM model susceptibility of wrongly judging case of high proficiency level and rapid answering, more suitable for measuring potential capability character of candidates.

Keywords: item response theory (IRT), computer adaptive testing, item response model, three-parameter logit model, Effort—moderated model, proficiency.

Introduction

In traditional paper-pen exams in a fixed time, each candidate has to answer a test from beginning to end, more frequently, higher capability candidates would feel items too simple, and complete test in advance so that excessive time was wasted; while lower capability candidate would think items too tough and have to guess answers, so testing result was hampered. Wise regarded that [1], prior to every testing, suppose the candidate will try the best to express knowledge he/she knows, and totally responded in testing results, in other words, candidate testing achievement derived from effort to pay in receiving test, if candidate did not exert all effort, candidate's actual proficiency level might be underestimated. Therefore, to measure candidate effort level, current researchers had developed many methods, e.g., employing post-test self-report scale or Person-fit statistics to estimate candidate effort level [1], later, Wise and Kong further proposed a new method: to measure candidate effort level according to item response time calculated from Computer Based Testing (CBT) [2]. The basis of theory was primarily from some studies of Schnipke and Scrams, they thought candidates often produce rapid answer response in time-limited testing, they generalized two kinds of candidate behavior, including (1) normal solution behavior: this type of candidates would actively look for right answer to test items; (2) rapid guess behavior: this type of candidates would produce rapid and random answer pattern [3,4,5,6,7].

Schnipke and Scrams also found that rapid guess behavior often happened in the end of time-limited testing, or when the candidate found test time was almost up, they would answer remaining test items soon. Wise and Kong assumed that, rapid guess behavior

implied such candidates lacked effort level. Therefore, to evaluate this hypothesis, they set an index of response time effort (RTE) to measure candidate overall testing effort level. RTE concept is mainly that, during testing, once a candidate answers a test item, there will be normal answer behavior or rapid guess behavior, behavior primarily depends on item response time the candidate takes to answer the item. Therefore, for a certain test item i , there will be a corresponding threshold time T_i , representing rapid guess behavior and normal answer behavior response time boundary. Suppose response time of a candidate j on test item i is RT_{ij} , then solution behavior index SB_{ij} can be defined as a binomial index, as shown in following formulas:

$$SB_{ij} = \begin{cases} 1 & \text{if } RT_{ij} \geq T_i \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad RTE_j = \frac{\sum_{i=1}^k SB_{ij}}{k}$$

where k is total of items in the test.

When a candidate showed normal answer behavior, the probability of candidate answering right would exhibit a monotonically increasing function with proficiency level, then three-parameter logistic model among item response model could be used to express probability of answering right. On the contrary, when a candidate was claimed as rapid guess behavior, candidate performance would be relatively not associated with proficiency level, showing a more flat response function instead, whose value was roughly equal to reciprocal of all option numbers for this item[8]. Combining item response functions derived from these two kinds of candidate behavior, there emerged a so-called Effort-moderated response model, EM response model in short, which can be expressed as follows:

$$P_i(\theta) = SB_{ij} * [c_i + (1 - c_i) * \left(\frac{1}{1 + e^{-Da_i(\theta - b_i)}} \right)] + (1 - SB_{ij}) * \frac{1}{d_i}$$

where, d_i is number of option of the i^{th} item, D is preset as a constant, 1.702, b_i is difficulty parameter of item i , a_i is discrimination parameter of item i , c_i is pseudo-chance parameter of item i , SB_{ij} is normal answer index of a candidate j on test item i .

However, in actual test, some candidates with high proficiency level would most probably infer the right answer to this item in predetermined threshold time, according to Effort-moderated response model, this type of candidates would be categorized as the rapid guess behavior, thus greatly underestimating right answer probability of such candidates, and might influencing adaptive level of following items; on the contrary, if a candidate did not know how to answer the item indeed, as answering time exceeds predetermined threshold time, he/she made guess behavior at the end, this type of candidates would be categorized into normal answer behavior and substituted to Three-parameter logistic model for capability calculation. Though Three-parameter logistic model itself had considered guess level, it still would somewhat overestimate probability of the candidate answering the item right. Based on above considerations, how to make up disadvantage of original Effort-moderated model, so that behavior of tested candidate can be more accurately determined, and proficiency level estimation can obtain a more reasonable value, which thus became the primary motive of this study.

1. Materials and Methods

1.1 Research Steps

In this paper, adaptive choice of test items was done according to individual capability scope of test candidates. Assessment was terminated by automatic determination after a complete cycle test. At the start of assessment, randomly select an item from intermediate difficulty region in database as the test starting point, though this item complies with difficulty parameter of -0.5~0.5 according to above analysis, discrimination parameter of

less than 0.25. After candidate answered, candidate capability was estimated by means of maximal likelihood capability estimation, if ending criterion preset for this system was satisfied, e.g., standard deviation difference is less than a preset value or prescribed testing item total was arrived, then stop testing; if testing ending condition was not reached, then adopt one-point maximal message approach as strategy of adaptive selection of items, according to candidate answering response, repeatedly estimate candidate capability, and verify ending criterion, until testing ended.

1.2 IEM model

Suppose the candidate proficiency level achieves to a certain superior level, it's very likely to have rapid answering behavior. Therefore, we could define rapid answer index A_{ij} as determinant whether a candidate was categorized to high proficiency level and generate rapid answer behavior. If proficiency level of candidate j is θ , rapid answer capability threshold on item i is PT_i , then rapid answer index A_{ij} can be defined in following formula:

$$A_{ij} = \begin{cases} 1, & \text{If } \theta \geq PT_i \\ 0, & \text{otherwise} \end{cases}$$

Hence, IEM model is defined in the following equation:

$$P_i(\theta) = (A_{ij} + SB_{ij} - A_{ij}SB_{ij}) * [c_j + (1 - c_j) * \left(\frac{1}{1 + e^{-Da_j(\theta - b_j)}} \right)] + (1 - A_{ij})(1 - SB_{ij}) * \frac{1}{d_i}$$

In this model, due to various combinations of A_{ij} and SB_{ij} indices, there would be four types to categorize candidate level and answer model analysis, hence, three combinations would use Three-parameter logistic model to calculate probability, and the fourth combination would use random guess probability model to analyze. Combinations were as follows: (1) when the candidate proficiency level higher than threshold with rapid answer response ($A_{ij}=1, SB_{ij}=0$), (2) the candidate with proficiency level above threshold and normally answered in a certain time ($A_{ij}=1, SB_{ij}=1$), (3) the candidate with proficiency level below threshold and normally answered in a certain time ($A_{ij}=0, SB_{ij}=1$), in the above three combinations, Three-parameter logistic model would be used to calculate probability of answering the item right; the last case was that when candidate proficiency level was below threshold and the answer was rapidly responded ($A_{ij}=0, SB_{ij}=0$), then the reciprocal of option numbers shall be the probability of answering the item right.

IEM model was mainly to solve the problem in actual testing that, if some high proficiency level candidates could infer or observe answer to the item right in predetermined time, then according to Effort-moderated response model, this type of candidates shall be categorized into rapid guess behavior, greatly underestimating their proficiency level and right answer probability, and further influencing later item adaptivity. In simulation study, it could be proved that using IEM model could solve this problem well, candidate capability estimation and item making strategy became more accurate, too. In this study, time and proficiency level threshold were used to distinguish normal answer behavior and rapid guess behavior, and the appropriate thresholds of IEM model is originated from. Wise[1].

1.3 Capability estimation

After candidate provides answer response, capability estimation method shall be relied on to calculate potential capability of candidate at that time, and from capability values estimated, select items adapting to candidate. In this paper, Maximal Likelihood Estimation (MLE) was employed as capability estimation method and the details of MLE procedures can be referred to Wikipedia.

1.4 Answer response simulation

To simulate answer response of candidate answering item, a set of logic process flow is needed, in other words, candidate answer response must be simulated according to its true capability, in case unreasonable data are estimated, hampering capability estimation accuracy. For example, a candidate whose true capability is -2.5 shall not be estimated all right answer response. Answer response simulation flow was briefed below.

Firstly, generate a R_j from uniform distribution of $[0,1]$, and use the model to calculate probability $P_j(\theta)$ of candidate answering the item right, if $P_j(\theta) \geq R_j$ then determine that the candidate answers it right, response is set as 1; if $P_j(\theta) < R_j$ then determine that the candidate answers it wrong, response is set as 0.

It's known from above that, if $S_j(\theta)$ is simulation answer response of candidate with capability θ to item j , then $S_j(\theta)$ is defined as follows:

$$S_j(\theta) = \begin{cases} 1, & \text{if } R_j \leq P_j(\theta) \\ 0, & \text{otherwise} \end{cases}$$

Where R_j depends on $U(0, 1)$, $U(0, 1)$ denotes uniform probability distribution between 0 and 1, $P_j(\theta)$ denotes probability of candidate with capability θ answering item j right.

1.5 Item library parameter simulation

To calculate $P_j(\theta)$ for IEM model, item parameter, A_{ij} and SB_{ij} shall be defined in advance. Item parameters adopted in this paper were obtained from joint maximum likelihood (JMLE) calculation of original candidate response data provided by National Middle School Basic Learning Test Steering Committee in Republic of China; A_{ij} is so taken, for first 50% of candidate capability distribution, $A_{ij}=1$, otherwise, $A_{ij}=0$; then SB_{ij} is determined to be 0 or 1 according to $U(0, 1)$.

Item selection method was to calculate item message amount of all untested items in item database, substitute capability value estimated of candidate to item message function, and select the maximal message amount item as next test item, or one-point maximal message method.

1.6 The evaluation of capability value

In simulation testing, after defining candidate response type, model can be chosen to estimate capability. Difference between estimated capability value and true capability value can be processed with root mean squared error (RMSE) as basis of later evaluation. The smaller the RMSE, the more approximate the estimated capability is to true capability, or more accurate the estimation is. If test sample number is N , candidate true capability value is θ_t , estimated capability in this system is θ_θ , then RMSE is defined below:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (\theta_t - \theta_\theta)^2}$$

2 Results and Discussions

Experimental data in this study were mainly obtained according to previous description of simulation study flow, calculate RMSE values in respect to IEM model, Effort-moderated model and Three-parameter logistic model for the same response category, and interpret its implication. If defining capability range at -3.0~3.0, with 0.1 as capability interval, then candidate capability ascending sequence is -3.0, -2.9, -2.8, ..., 0, 0.1, ..., 2.8, 2.9, 3.0, 61 groups in total, run 1000 person/times simulation tests for each varied candidate capability, each testing length is fixed to 30 items. At the initiate of test, randomly select one item from intermediate difficulty region in item database as testing starting point, provided that its difficulty parameter must be in -0.5~0.5, discrimination parameter shall be less than 0.25. According to simulation response category, use maximal likelihood capability estimation

method to estimate candidate capability, during the estimation, use three item response models separately and record estimated capability by each model. If fixed testing item total is arrived, then stop testing; if testing has to go on, then use one-point maximal message method as adaptive selection strategy. Continuing to estimate capability according to simulation answer response, till end of testing is achieved.

Table 1 showed RMSE of capability values of three models estimated according to specific simulation response category, specific simulation response category was to highlight high proficiency level and rapid answer, and random guess behavior due to low proficiency level, increase frequency of these two behaviors in item parameters to about 90%, in other words, if candidate capability exceeds high point cluster threshold, then if nearer to 3, it will be more likely to set A_{ij} as 1, and SB_{ij} as 0. If candidate capability is below low point cluster threshold, then if nearer to -3, it will be more likely to set A_{ij} as 0, and SB_{ij} as 0. From Table 1 that, in testing with fixed length of 30 items, if comparing high point cluster RMSE of IEM model and EM model, we could find that IEM model had significantly lower RMSE, in other words, in case of rapid answer in high point cluster, IEM model could accurately estimate candidate capability, unlike EM model, which wrongly determined candidates with high proficiency level and rapid answer capability wrongly as random guess, thus resulted in greater RMSE error. Therefore, as to candidates with high proficiency level and rapid answer capability, IEM model could accurately measure proficiency level. If comparing low point cluster RMSE of IEM model and 3PL model, it could be found that IEM model RMSE was similar to that of 3PL model, indicating that, when there was random guess behavior due to low candidate proficiency level, using 3PL model to estimate its capability, candidate right answering probability is $c_j + (1 - c_j) * \left(\frac{1}{1 + \theta^{-Da_j} (g - b_j)} \right)$, which might be slightly higher than reciprocal of item option numbers in IEM model. Therefore, as low proficiency level candidates have lower capability, 3PL model would slightly higher than capability estimated, using IEM model can accurately estimate capability.

	Overall RMSE	High point cluster RMSE	Low point cluster RMSE	Ending condition
IEM model	0.2882	0.2675	0.0207	Fixed length method
3PL model	0.2904	0.2697	0.0206	Fixed length method
EM model	0.8175	0.7970	0.0205	Fixed length method

Table 1 three model distribution of capability values of three models to specific item parameter estimation.

References

- [1] Steven L. Wise, DeMars and Christine E. (2006). An Application of Item Response Time: The Effort-Moderated IRT Model, *Journal of Educational Measurement*, v43 n1, p19-38.
- [2] Wise, S. L., & Kong, X(in press) . (2005). Response time effort: A new measure of examinee motivation in computer-cased tests. *Applied Measurement in Education*, vol. 18, no. 2, p 163-183.
- [3] Schnipke, D. L.(1995). Assessing speededness in computer-based tests using item response times. Paper presented at the annual meeting of the National Council on *Measurement in Education*. San Francisco, CA, April.
- [4] Schnipke, D. L. (1996). How contaminated by guessing are item-parameter estimates and what can be done about it? . Paper presented at the annual meeting of the National Council on *Measurement in Education*. New York, NY, April.
- [5] Schnipke, D. L. (1999). The influence of speededness on item-parameter estimation (*Computerized Testing Report No. 96-07*) . Princeton, NJ: Law School Admission Council.
- [6] Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, 34, 213-232.
- [7] Schnipke, D. L., & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In Mills, C. N., Potenza, M.T., Fremer, J. J., & Ward, W. C. (Eds.) . *Computer-based testing: Building the foundation for future assessments*. Mahwah, NJ: Lawrence Erlbaum Associates.
- [8] Wise, S. L. (2004). An investigation of the differential effort received by items on low-stakes, computer-based tests. Manuscript under review.