Paper PO-23

A Coding Practice for Preparing Adaptive Multistage Testing

Yung-chen Hsu, GED Testing Service, LLC, Washington, DC

ABSTRACT

The purpose of this paper is to present a simulation study of a coding practice for preparing adaptive multistage testing (MST) designs for a credentialing testing program in the coming years. MST is an adaptive test administration method in which a test form is tailored as a sequence of pre-constructed modules at item set level. At each adaptation point a module is selected to match the proficiency estimate of the examinee based on cumulative performance on previously administered modules. For some testing programs, MST is considered a better fit in their future test development because the test delivery model offers a balanced tradeoff and a promising amelioration between the computerized adaptive tests and the traditional linear fixed-length tests. In the simulation, a macro is developed to estimate the proficiency scores based on item response theory. The algorithm is implemented with PROC IML using Newton-Raphson method. To assess the classification consistency and decision accuracy for examinees, kappa coefficients from PROC FREQ and additional consistency measures are computed to more fully characterize the extent of the agreement. Practical policy questions and test development considerations are also discussed.

INTRODUCTION

For many credentialing testing programs, using computers to administer exams is a trend in the coming years. There are good practical reasons why adopting a computer-based test administration is preferred, which include automated scoring and fast reporting, flexible exam schedules and locations, and potentially higher efficiency and more precise proficiency estimation of examinees through computer-based or adaptive testing.

Several innovative test delivery models were considered in practice, such as computerized fixed testing (CFT), item-level computer-adaptive testing (CAT), and multistage testing (MST) (Henrickson, 2007; Jodoin, Zenisky, and Hambleton, 2006). A CFT is analogous to the fixed-item paper-and-pencil test (PPT) but with more modern varieties that can be administered. For example, different examinees may take different forms of the test or receive the items in different orders. In contrast, CAT adapts the difficulty of the test and presents each new item based on the proficiency estimate of an examinee's performance on previous items. CAT generally uses much fewer test items than CFT does and has the advantage of offering improved efficiency in estimating examinee's proficiency level. However, there are potential psychometric issues and practical shortcomings found in the past CAT practices, such as item exposure control and content balancing. Besides, data management effort and deployment cost are business and financial concerns for administering the test in operational environment. To balance the tradeoff between CFT and CAT, MST was proposed. MST is a test administration method closely related to CAT but has a test adaptation at the item set level instead. For some examination programs MST is considered as a better delivery model in the test development. MST may help ameliorate the problems encountered in a traditional CAT yet still offers better testing efficiency than PPT or CFT.

Depending on the test nature and practical policy of the testing program, there are practical test development considerations related to the design and implementation. This study is a coding practice using simulated data in preparing information for decision makers of a credentialing testing program in the future.

ADAPTIVE MULTISTAGE TESTS

Figure 1 depicts the generalized procedure of adaptive multistage test design. A test form consists of a series of test modules and each test taker would potentially take a different set of modules that is best targeted to the individual's ability. MST starts with an initial test module for all examinees. The initial module commonly contains items with moderate difficulty at the median proficiency of the intended group or a broad range of difficulty values. With the examinee's performance on the initial module, a proficiency score can be estimated. The proficiency estimate is then used to select the next module that matches the examinee's proficiency level. Normally, the accumulated performance was used to estimate the proficiency for deciding a module with narrow and more focused difficulty in each round until the test ends.

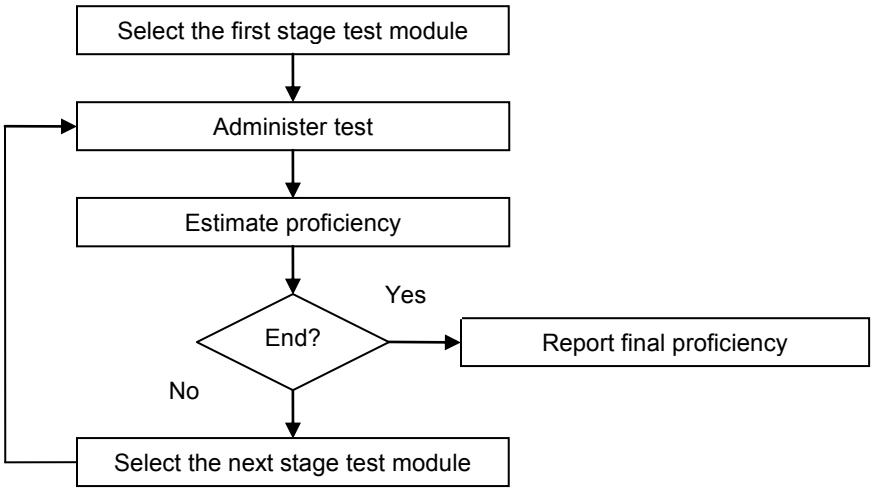

Figure 1. MST procedure

SIMULATION

A simulation using the two-stage test design was conducted to demonstrate the procedure for the preparation work of the test development. Rasch model is used in this simulation, which is the simplest item response theory (IRT) model for dichotomous item having only one parameter for the examinee and one for the item that generically referred to as a threshold. Mathematically, Rasch model can be expressed as

$$P(u_{ij} = 1) = \frac{1}{1 + \exp(b_i - \theta_j)}$$

representing the probability of answering a particular dichotomously scored item correctly given the proficiency level of a test taker, where b_i is the difficulty of item i while θ_j is the ability of person j.

The steps of conducting the simulation study are outlined as follows:

- Data preparation: Simulate true proficiency scores (θ_t), item parameters (b), item responses (u), true group ID, a single stage module, and both first and second stage modules.
- Proficiency estimation: Use u and b from the first stage to estimate proficiency score θ_1. Based on θ_1 to assign second stage modules. Combine data from both stages and estimate the final proficiency score θ_{12}. Also estimate single stage proficiency scores θ_0 for comparison.
- Evaluation: Calculate psychometric properties and related statistics for evaluation.

DATA PREPARATION

To simulate true proficiency scores for 3,000 test candidates, θ_t are generated from a normal distribution $N(0,3)$ with predefined upper and lower bounds. We assume that the test will be used to classify the candidates into three groups: A, B, and C (e.g., pass advanced, pass, and fail). The candidates are divided into three groups with 1,000 each according to the true proficiency scores θ_t. Three sets of item parameter b are also generated from a normal distribution with different mean and bounds. The three sets are combined form a pool of 60 items. Then, the responses u are generated from θ_t and b using Rasch model. One first stage module, which contains 30 items, and three second stage modules with 20 items each are assigned. A 50-item single-stage module is also assigned for comparison. The first stage module is designed to contain a broad range of difficulty values while the second stage modules has more items with difficulty located near the average proficiency scores of the respective group. Figures 2 and 3 illustrate the test information curves of the first and second modules, respectively. The test information is simply the sum over items of the amount of item information. Namely,

$$I(\theta) = \sum I_i(\theta).$$

The item information function is defined as

$$I_i(\theta) = \frac{[P_i'(\theta)]^2}{P_i(\theta)Q_i(\theta)},$$

where $Q_i = 1 - P_i$, and $P_i'(\theta) = \partial P_i(\theta)/\partial\theta$. For Rasch model, the expression is simply

$$I_i(\theta) = P_i(\theta)Q_i(\theta).$$

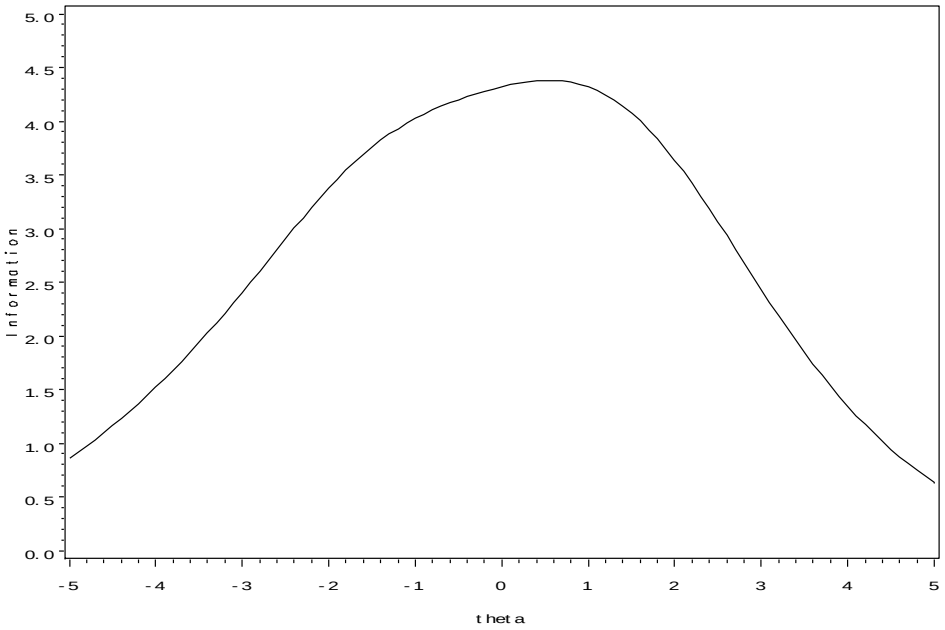

Figure 2. Test information curve of the first stage module

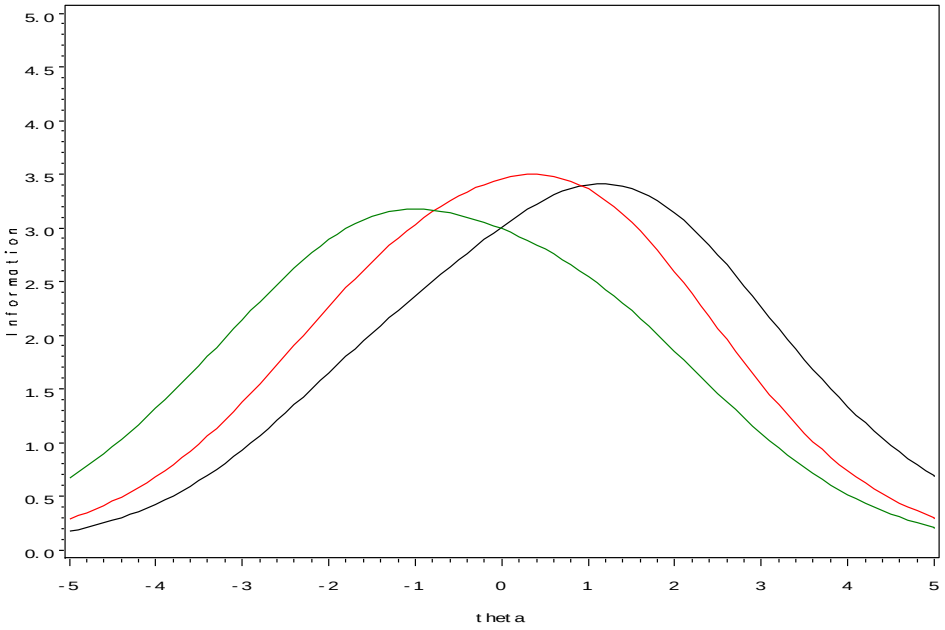

Figure 3. Test information curves of the second stage modules

PROFICIENCY ESTIMATION

The measure of the proficiency or ability of a given examinee is the maximum likelihood estimate based on the responses to the items and the values of the parameters of the items. For a test module with N item, $u_i \in \{0,1\}$ refers to a test taker's response to item i, which is scored dichotomously. Under the assumption of local independence, the probability of the vector of item response $U = (u_1, u_2, \cdots, u_N)$ is given by the likelihood function

$$\Pr(U|\theta) = \prod_i P_i^{u_i} Q_i^{1-u_i} ,$$

where P_i is the Rasch function, $Q_i = 1 - P_i$, and θ is the ability of the test taker. The derivatives of the log-likelihood function with respect to the test taker is

$$\frac{\partial L}{\partial \theta} = \sum u_i P_i^{-1} \frac{\partial P_i}{\partial \theta} + \sum (1-u_i) Q_i^{-1} \frac{\partial Q_i}{\partial \theta},$$

where L is the natural logarithm of the likelihood function $\Pr$. For Rasch model, we have

$$\frac{\partial P_i}{\partial \theta} = P_i Q_i \text{ and } \frac{\partial Q_i}{\partial \theta} = -P_i Q_i .$$

Then

$$\frac{\partial L}{\partial \theta} = \sum (u_i - P_i) \text{ and } \frac{\partial^2 L}{\partial \theta^2} = -\sum P_i Q_i .$$

Using a Taylor series expansion to solve the likelihood equations, we have

$$\frac{\partial L(\theta)}{\partial \theta} = \frac{\partial L(\theta_0)}{\partial \theta} + (\theta - \theta_0) \frac{\partial^2 L(\theta_0)}{\partial \theta^2} + \ldots = 0$$

where θ_0 can be viewed as a trial value for the root of θ at the n th step. The approximate value of the next step θ_{n+1} can be derived from

$$\Delta\theta = \theta_{n+1} - \theta_n = \frac{\sum (u_i - P_i)}{\sum P_i Q_i} .$$

with second-order approximation. The above iterative scheme is known as the *Newton-Raphson* method and the process must be repeated until $\Delta\theta$ become sufficiently small.

A SAS/IML® module, which implemented the Newton-Raphson method, is used to simplify the task. The following statements show a macro that calls the IML module to estimate the ability for every test taker in an iteration loop. The initial trial values are all set to be zero in the macro. To improve the efficiency, they can first be replaced and estimated by using the total score or other means.

```
%macro rb2tRasch(  /* Ability estimation        */
  dsR=,            /* item response            */
  dsP=,            /* Item parameter           */
  dsT=             /* Ability                  */
);

proc iml;
  nMaxIter=30;    *max iteration number;
  minDelta=0.01;
  ubTheta=5;      *theta upper bound;
  lbTheta=-5;     *theta lower bound;

  use &dsR; read all var _num_ into r;
  use &dsP; read all var _num_ into b;
```

```
  nTakers=nrow(r);
  nItems=ncol(r);
  nItems1=nrow(b);

  * Error check;
  if nItems^=nItems1 then do;
    print "ERROR: Inconsistent inputs."; stop;
  end;

  * Newton-Raphson equation iteration loop;
  start rascht(t0,pb,r,mxit,ubt,lbt,mnd);
    t=t0;
    nit=1;
    ni=nrow(pb);
    do while(nit<=mxit);
      snum=0.0; sdem=0.0;
      do i=1 to ni;
        p=1.0/(1.0+exp(pb[i]-t));
        w=p*(1.0-p); v=r[i]-p;
        snum=snum+v; sdem=sdem+w;
      end;
      dta=snum/sdem;

      * Check convergence and set bounds;
      if abs(dta)<mnd then nit=mxit;
      else if dta>ubt then delta=ubt;
      else if dta<lbt then delta=lbt;

      * Update;
      t=t+dta;
      nit=nit+1;
    end;
    return (t);
  finish rascht;

  * Initial estimate.
  t0=j(nTakers,1,0);

  * Loop through every test taker;
  theta=j(nTakers,1,0);
  do j=1 to nTakers;
    theta[j]=rascht(t0[j],b,r[j,],nMaxIter,ubTheta,lbTheta,minDelta);
  end;
  create &dsT from theta[colname='theta']; append from theta; close theta;
  quit;
run;
%mend rb2tRasch;
```

EVALUATION

The correlation matrix of the true proficiency scores, estimated scores from the two-stage (30 and 20 items) test and from the single stage test is provided in Table 1 by using PROC CORR procedure. The correlation between the true score and the two-stage proficiency estimates is higher than the correction between the true score and the single-stage proficiency estimates.

Table 1. Correlation matrix of true scores, two-stage, and single stage proficiency estimates

	True score	Two-stage
Two-stage	0.85134	
Single-stage	0.70285	0.79453

For most credentialing testing programs the decision accuracy for classifying candidates is crucial. We assumed that both A and B groups are collapsed as Pass, and C group is Fail. Then the Cohen's kappa coefficient, which provides a measure of agreement, can be obtained from PROC FREQ procedure with TEST KAPPA option as

$$\kappa = \frac{P_o - P_c}{1 - P_c}$$

where P_o is observed agreement and P_c is chance agreement. The values of kappa range from -1 to +1. However, negative kappa is unusual in practice as the observed agreement is less than change agreement. A number of studies (Sim and Wright, 2005; Viera and Garrett, 2005) show that there are other factors can influence the magnitude of kappa and suggested reporting additional indices for providing a clear picture, such as prevalence index and bias index. Both of them are included in Table 2 although low kappa and high prevalence are very rare in most well designed educational assessment program. The decision accuracy of the two-stage case is slightly higher but not significant in the simulation. The mean and standard deviation of the difference to the true score are also provided even though accurate proficiency estimates are less critical for most credentialing tests. The results show that two-stage design yields more accurate estimates.

Table 2. Cohen's kappa, Prevalence index, and Bias index

	Cohen kappa	Prevalence index	Bias index	Mean	Standard deviation
True/Two-stage	0.8516	0.3257	0.0077	0.3421	1.0768
True/Single-stage	0.8475	0.3357	0.0023	0.4862	2.1657

CONCLUSION

This simulation study is a coding practice of preliminary preparation work for a credentialing testing program in the coming years. MST is being considered and is expected to have some distinct advantages over conventional fixed-length testing. In the test development, there are many issues to resolve. In order to provide information for decision making, parameters and data will be adjusted accordingly when more information, such as data collected from field testing, item characteristics in the future item bank, and data derived from previous tests, become available. This paper illustrates the procedure using SAS in preparing information for making decision and outlines some steps for use in the development.

REFERENCES

Hendrickson, A. (2007). An NCME Instructional model on multistage testing. *Educational Measurement: Issues and Practice,* 26(2), 44-52.

Jodoin, M.G., Zenisky, A., and Hambleton, R.K. (2006). Comparison of the psychometric properties of several computer-based test designs for credentialing exams with multiple purposes. *Applied Measurement in Education*, 19(3), 203-220.

Sim, J. and Wright, C.C. (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical therapy*, 85(3), 257-268.

Viera, A.J. and Garrett, J. M. (2005). Understanding interobserver agreement: The kappa statistic. *Family Medicine*, 37(5), 360-363.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Yung-chen Hsu
GED Testing Service, LLC
One Dupont Circle NW
Washington, DC 20003
Work Phone: 202-939-9717
E-mail: yung-chen.hsu@GEDtestingservice.com
Web: www.gedtestingservice.com