Paper 241

Generating Item Responses Based on Multidimensional Item Response Theory

Jeffrey D. Kromrey
Cynthia G. Parshall
Walter M. Chason
Qing Yi

University of South Florida

ABSTRACT

The purpose of this paper is to demonstrate code written in SAS/IML® software that generates examinees' test responses (0/1s) based on a multidimensional item response theory (MIRT) model. This program reads in a file of calibrated item parameters from the NOHARM computer program (Fraser & McDonald, 1986) and generates normally distributed random variables to represent examinees' ability levels on each dimension.

The SAS/IML program calculates the probability of an examinee obtaining a correct response based on the MIRT model, then compares this probability with a uniform random number to decide the examinee's item response. If the probability is larger than the random number, the examinee is credited a correct response (i.e., an item score of 1), otherwise, a zero. The program allows control of the number of samples, the number of examinees, and the number of items for which item responses are generated.

INTRODUCTION

Item response theory (IRT; Lord, 1952, 1953a, 1953b) applies a set of mathematical models to indicate the interaction between an examinee's ability (θ) or a composite of abilities and the characteristics of items in a test. In IRT models, $\hat{\theta}$ is used to denote an examinee's estimated level of the latent trait, ability, or skill that is measured by test items. Many different types of models have been developed in IRT (e.g., van der Linden & Hambleton, 1996). In this presentation, however, attention is focused on three-parameter models for dichotomously scored items (i.e., correct/not correct; 0/1).

In IRT, as an examinee's ability (θ) increases so does the probability of answering an item correctly. The probability of an examinee answering an item correctly in the three-parameter logistic IRT model can be defined as

$$P_i(\theta_j)=c_i+\frac{1-c_i}{1+e^{-Da_i(\theta_j-b_i)}}$$

where

- e is the base of the natural logarithms and equals 2.71828…,
- i indexes test item ($i=1,2,3,\ldots,n$),
- j indexes examinee, and $j=1,2,3,\ldots,N$,
- a_i is the item discrimination index for item i, that is proportional to the slope of the item response function at the point $\theta_j=b_i$,
- b_i is the item difficulty index for item i, that is the point on the ability scale at which an examinee has $(1+c)/2$ probability of answering item i correctly,
- c_i is the lower asymptote parameter of the item response function for item i, that represents the probability of examinees with very low ability correctly answering the item,
- θ_j represents the ability of examinee j,
- $P_i(\theta_j)$ is the probability of examinee j with ability level θ answering item i correctly, and
- D is a scaling factor that equals 1.702.

IRT includes a group of assumptions about the data to which the models apply (Hambleton, Swaminathan, & Rogers, 1991). One assumption is called the assumption of unidimensionality, which means that only one ability or one composite of multiple abilities is measured by a test. However, many educational

and psychological tests measure several latent traits rather than a single one (Reckase, Ackerman, & Carlson, 1988; Traub, 1983) and the extent to which individual items reflect each trait can vary from item to item (Ackerman, 1994b). For example, a simple mathematics story problem may require both reading and mathematics skills to provide a correct response. Examinees may bring a variety of cognitive skills to a testing situation, some of which may be used during the test and some not. Miller and Hirsch (1992) indicated that substantial measurement problems may arise if a unidimensional item analysis procedure is used with multidimensional items. For example, problems can occur in the process of constructing a test using classical test theory procedures, or when the statistics provide no indication of what abilities are being measured by items or how well each ability is measured.

Therefore, researchers have been advised to use MIRT when the unidimensionality assumption is violated (Reckase, 1985; Ackerman, 1994a). The MIRT models do not require the assumption of unidimensionality.

The probability of a correct response to item i in an k-dimensional logistic model (Reckase, 1985) can be expressed as

$$P(u_i=1 \mid \theta_j)=c_i+(1-c_i)\frac{\exp[1.702\mathbf{a}_i'\theta_j+d_i]}{1.0+\exp[1.702\mathbf{a}_i'\theta_j+d_i]}$$

where

u_i is examinee's score (0/1) on item i $(i=1,2,3,\ldots,n)$,

$\mathbf{a}_i$ is the vector of item discrimination parameters $(a_{ik}=a_{i1},a_{i2},a_{i3},\ldots,a_{im})$ for item i in k dimensions $(k=1,2,,3,\ldots,m)$,

d_i is the scalar difficulty parameter for item i, negative d values represent difficult items, and positive values represent easy items,

c_i is the scalar lower asymptote parameter for item i,

θ_j is the vector of $\hat{\theta}$ for person j $(j=1,2,3,\ldots,N)$, and

$P(u_i=1 \mid \theta_j)$ is the probability of an examinee j correctly answering item i.

In this model, there is an item discrimination parameter for each dimension of the model but only one overall item difficulty parameter. The components in the function are additive, thus, being low on one latent trait can be compensated for by being high on another trait.

NOHARM (Fraser & McDonald, 1986) and TESTFACT (Wilson, Wood, & Gibbons, 1984) are two of the computer programs that estimate parameters for the MIRT model. NOHARM (Fraser & McDonald, 1986) is a computer program that fits the normal ogive model by a least-squares procedure and will estimate a_{ik} and d_i parameters in the MIRT model. NOHARM (Fraser & McDonald, 1986) does not estimate the c_i parameters but requires values to be input and treated as fixed. Usually the c_i parameters are estimated from a unidimensional analysis using a computer program such as BILOG (Mislevy & Bock, 1990). Miller and Hirsch (1992) indicated that asymptotically the c_i values are the same for models of any number of dimensions. The original NOHARM (Fraser & McDonald, 1986) program can handle as many as six dimensions of item parameters in a multidimensional space.

Previous research has indicated that simulated data based on a MIRT model are more similar to real test data than are data generated by other approaches (Davey, Nering, & Thompson, 1997; Parshall, Kromrey, Chason, & Yi, 1997). Recently, researchers have used MIRT as the basis of data simulations (e.g., Parshall, Davey, & Nering, 1998; Yi & Nering, 1998a, 1998b). The SAS code in this presentation demonstrates how to use SAS/IML to simulate data according to a MIRT model.

GENMIRT.SAS

The program GENMIRT.SAS uses SAS/IML to simulate examinees' responses according to the MIRT model. The program requires, as input, MIRT parameters for a set of test items. These parameters may be obtained from a MIRT calibration program, such as NOHARM (Fraser & MacDonald, 1986). The output from the program is an ASCII file of item scores (0/1s), representing correct and incorrect

responses to each test item. The output file is a series of N *X* K matrices, in which N is the number of examinees simulated, and K is the number of items on the simulated test. The item score matrices are augmented with examinee identification numbers and the examinee ability level (θ) for each dimension.

The program GENMIRT.SAS operates in six major steps, as follows:

1. *Read MIRT item parameters*. As written, the MIRT parameters are read from an external file, into a SAS data step, then passed to SAS/IML.

2. *Establish the number of samples and number of examinees to generate*. The two nested do loops, (DO REP = 1 to 100, and DO I = 1 to 1000) establish, respectively, the number of samples to generate and the number of examinees in each sample for whom responses will be simulated. Simply changing the maximum values of REP and I in these two loops changes the number of samples or number of examinees to be simulated.

3. *Simulate examinee ability on each dimension*. Generate six random numbers from an NID(0,1) distribution. These values are used as the examinees' true ability levels on the six MIRT dimensions.

4. *Generate a uniform random number for each examinee and for each item on the test*. To simulate the probabilistic nature of test item responses, a uniform random number (U), on the 0 to 1 interval, is compared to the calculated probability of a correct response for each item (P_i). If P_i> U then the examinee is credited with a correct response to the item (receiving an item score of 1). Conversely, if P_i<=U then the examinee obtains an incorrect response to the item (receiving an item score of 0).

5. *Calculate a vector of item scores for each examinee*. The subroutine IRTSCORE is used to calculate each examinees' probability of correct response to each item. The inputs to this subroutine are the number of test items for the simulation (the scalar quantity NITEMS), the 1 *X* 6 vector of examinee ability parameters (SIMULEES), an examinee identification number (IDN2), the 1 *X* NITEMS vector of uniform random numbers that are compared to the probabilities of correct responses for the set of items (RRV), and the vectors of MIRT item parameters (POPA, POPB, and POPC). For each item, the probability of a correct response is calculated using the PROBNORM function, and the probability is compared to the value of the uniform random number. The subroutine returns a vector of 1s and 0s that represent the examinee responses to the set of items (SCORE).

6. *Create the output file*. The elements of the vectors SIMULEES and SCORE are placed into scalars so that the FILE and PUT statements will write them to an ASCII file.

PROGRAM CODE

```
options ls=182 ps=32767 pageno=1 formdlim='-';

proc printto print='c:\i500a.raw';

* +-----------------------------------------------------------+
     GENMIRT.SAS
     Generate a file of item scores (0,1) based on
     six-factor MIRT model.
  +-----------------------------------------------------------+;

   data params;

* +-----------------------------------------------------------+
     This is a file of known item parameters,
     separated by at least one blank and including
     an item number on each record.
  +-----------------------------------------------------------+;

 infile 'a:\IN80.PRS' lrecl=124 missover;
 input itemnum a1 a2 a3 a4 a5 a6 b c;

proc iml;



* +-----------------------------------------------------------+
     Define the subroutine to analyze each
     examinee response vector.
  +-----------------------------------------------------------+;

start irtscore (nitems, simulees, idn2, rrv, popa,
popb, popc, score);

 factnorm=PROBNORM(popb+(popa*simulees));

 *+-----------------------------------------------------------+
     The following yields a vector of probabilities
     of correct responses on each item (pi).
  +-----------------------------------------------------------+;

 Pi = (popc + ((1 - popc) # factnorm))' ;
```

```
 *+-------------------------------------------------------+
     The following yields the score vector (1,0)
  +-------------------------------------------------------+;

 score = Pi>rrv;
finish;

  use params;

* +-------------------------------------------------------+
     Reading in the vectors of item parameters
  +-------------------------------------------------------+;

  read all var {a1 a2 a3 a4 a5 a6} into popa;
  read all var {b} into popb;
  read all var {c} into popc;

  nitms=nrow(popa);

* +-------------------------------------------------------+
     This loop generates six theta values for each
     examinee and a set of NITMS random
     numbers.
  +-------------------------------------------------------+;

DO REP = 1 TO 100;
 DO I = 1 to 1000;
  seed1=round(100000000*ranuni(0));
  idn2 = i;
* +-------------------------------------------------------+
     Generation of theta values from N(0,1)
     distribution
  +-------------------------------------------------------+;

   sim1 = rannor(seed1);
   sim2 = rannor(seed1);
   sim3 = rannor(seed1);
   sim4 = rannor(seed1);
   sim5 = rannor(seed1);
   sim6 = rannor(seed1);
   simulees = sim1//sim2//sim3//sim4//sim5//sim6;

* +-------------------------------------------------------+
     Generation of uniform random numbers for
     each person and each test item. These are
     used to determine item response
     correctness.
  +-------------------------------------------------------+;

   rrv = J(1,nitms,0);
   do k = 1 to nitms;
    rrv[1,k] = RANUNI(seed1);
   end;

* +--------------------------------------+
     Call the scoring subroutine
  +--------------------------------------+;

  run irtscore (nitms, simulees, idn2, rrv, popa,
popb, popc, score);

* +-------------------------------------------------------+
     Create variables for the output data file
  +-------------------------------------------------------+;

    idnum = idn2[1,1];
    thet1 = simulees[1,1];
    thet2 = simulees[2,1];
    thet3 = simulees[3,1];
    thet4 = simulees[4,1];
    thet5 = simulees[5,1];
    thet6 = simulees[6,1];
    itm1  = score[1,1];
    itm2  = score[1,2];
    itm3  = score[1,3];
    itm4  = score[1,4];
    itm5  = score[1,5];

    [ etc. for each item]

    itm79 = score[1,79];
    itm80 = score[1,80];
  file print  ;
  put @1 idnum 4. @6 thet1 12.8
       @20 itm1 1.
       @21 itm2 1.
       @22 itm3 1.
       @23 itm4 1.

    [ etc. for each item]

       @98 itm79 1.
       @99 itm80 1.
       @110 thet2 12.8
       @125 thet3 12.8
       @140 thet4 12.8
       @155 thet5 12.8
       @170 thet6 12.8;
  end;
 end;
quit;
```

CONCLUSION

GENMIRT.SAS provides a simple vehicle for the simulation of realistic examinee test item responses. The data simulated by this program may be used for research on a variety of issues related to psychometrics, such as the accuracy and precision of methods to estimate examinee ability, strategies for test equating, phenomena

associated with computer adaptive testing algorithms, and techniques to detect differential item functioning.

REFERENCES

Ackerman, T. A. (1994a). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education, 7(4),* 255-278.

Ackerman, T. A. (1984b). Creating a test information profile for a two-dimensional latent space. *Applied Psychological Measurement, 18(3),* 257-275.

Davey, T., Nering, M., & Thompson, T. (1997, March). *Realistic simulation of item response data.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Fraser, C. & McDonald, R. (1986). *NOHARM II: A FORTRAN program for fitting unidimensional and multidimensional normal ogive models of latent trait theory.* Amidale, Australia: University of New England, Center for Behavioral Studies.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamental of item response theory.* Sage Publications.

Lord, F. M. (1952). A theory of test scores. *Psychometric Monograph,* 7.

Lord, F. M. (1953a). An application of confidence intervals and maximum likelihood to the estimation of an examinee's ability. *Psychometrika, 18,* 57-75.

Lord, F. M. (1953b). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13,* 517-548.

Miller, T. R. & Hirsch, T. M. (1992). Cluster analysis of angular data in applications of multidimensional item-response theory. *Applied Measurement in Education, 5(3),* 193-211.

Mislevy, R. J. & Bock, R. D. (1990). *BILOG3: Item analysis and test scoring with binary logistic models.* [Computer program]. Chicago, IN: Scientific Software.

Parshall, C. G., & Davey, T., & Nering, M. (1998, April). *Test development exposure control for adaptive testing.* In T. Miller (chair), Adaptive Testing Research at ACT. Symposium conducted at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Parshall, C. G., Kromrey, J. D., Chason, W. M., & Yi, Q. (1997, June). *Evaluation of parameter estimation under modified IRT models and small samples.* Paper presented at the annual meeting of the Psychometric Society, Gatlinburg, TN.

Reckase, M. D. (1985, April). *The difficulty of test items that measure more than one ability.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement, 25(3),* 193-203.

Traub, R. E. (1983). A priori considerations in choosing an item response model. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 57-70). Vancouver, BC: Educational Research Institute of British Columbia.

van der Linden, W. J. & Hambleton, R. K. (1996, Eds.). *Handbook of modern item response theory.* New York, NY: Spinger-Verlag.

Wilson, D., Wood, R., & Gibbons, R. (1984). *TESTFACT: Test scoring and full-information item factor analysis.* [Computer program]. Mooresville, IN: Scientific Software, Inc.

Yi, Q. & Nering, M. (1998a, April). *Nonmodel-fitting responses and robust ability estimation in a realistic CAT environment.* Paper presented at the annual meeting of American Educational Research Association, San Diego, CA.

Yi, Q. & Nering, M. (1998b, April). *The impact of nonmodel-fitting responses in a realistic CAT environment.* In M. Nering (chair), Innovations in person-fit research. Symposium conducted at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

CONTACT INFORMATION

The authors can be contacted at the University of South Florida, Department of Educational Measurement and Research, FAO 100U, 4202 East Fowler Ave., Tampa, FL 33620, by telephone (813) 974-3220, or Jeff can be contacted by e-mail: *kromrey@typhoon.coedu.usf.edu*