SAS® GLOBAL FORUM 2017

April 2 – 5 | Orlando, FL

Classification Decision Accuracy and Consistency in IRT by Using SAS/IML®

USERS PROGRAM

Classification Decision Accuracy and Consistency in IRT by Using SAS/IML®

Seohee Park[1], Kyungyong Kim[1], and Sung-Hyuck Lee[2]

[1]University of Iowa and [2]ACT

ABSTRACT

- Classification Decision (CD) has been predominantly used in educational and vocational situations such as admissions, selection, placement, and certification. This method needs to be accurate because the results have important implications for examinees' professional and academic futures.
- Classification Accuracy and Consistency statistics (CA/CC) are indices representing the precision of CD, and they need to be reported in order to affirm the validity of the CD.
- Classification Accuracy (CA) is referred to as the degree to which the classification of observed scores matches with the classification of true scores. Classification Consistency (CC) is defined as the degree to which examinees are classified in the same category when taking two parallel test forms (Lee, 2010).
- Under Item Response Theory (IRT), both Rudner (2001) and Lee (2010) have developed methods to calculate CA/CC.
- The purpose of this paper is to introduce a IRT Classification Accuracy and Consistency SAS/IML® program.
- Application of the program was conducted with testlet-based tests, which can be analyzed in 3 parameter Logistic model (3PL) and the Graded Response Model (GRM) and the Generalized Partial Credit Model (GPCM) for the testlet-as polytomous IRT model.

METHODS

- **Rudner method** : The probability ($\gamma_p(\widehat{\theta_j}|\theta_j)$) that an examinee's estimated theta is above a cut score and the probability ($\gamma_f(\widehat{\theta_j}|\theta_j)$) that the examinee's estimated theta is below the cut score are calculated. The cut score (θ_c) here is on the latent trait scale.

$$\gamma_p(\widehat{\theta_j}|\theta_j) = \int_{\theta_c}^{\infty} \frac{1}{\sqrt{2\sigma_{\widehat{\theta_j}}^2 \pi}} e^{-\frac{(\theta_j - \widehat{\theta_j})^2}{2\widehat{\theta_j}}} d\theta_j, \quad \gamma_f(\widehat{\theta_j}|\theta_j) = 1 - \gamma_1(\widehat{\theta_j}|\theta_j)$$

- **Lee method:** The probability (τ_p) that an examinee's estimated theta is above a cut score, and the probability (τ_f) that an examinee's estimated latent trait is below the cut score are calculated with the follow. The cut score (x_c) here is total score matric.

$$\tau_p = \sum_{x=x_c}^{X} \phi(x|\hat{\theta}), \qquad \tau_f = \sum_{x=0}^{x_c-1} \phi(x|\hat{\theta})$$

Where $\phi(x|\hat{\theta})$ is the conditional score distribution given theta calculated with Lord-Wingersky Method.

- **Classification Accuracy** : $\frac{\sum_j^N \{\gamma_p(\widehat{\theta_j}|\theta_j) * w_{pj} + \gamma_f(\widehat{\theta_j}|\theta_j) * w_{fj}\}}{N}$, $where\ j = examinee\ (1, 2, 3, \dots, N)$ where w_{pj} be 1 if $\widehat{\theta_j}$ is above the cut score, which indicates that the true latent trait is above the cut score, and w_{fj} be 0 if $\widehat{\theta_j}$ is below the cut score.
- **Classification Consistency** :$\frac{\sum_j^N \{\gamma_p(\widehat{\theta_j}|\theta_j) * \gamma_p(\widehat{\theta_j}|\theta_j) + \gamma_f(\widehat{\theta_j}|\theta_j) * \gamma_f(\widehat{\theta_j}|\theta_j)\}}{N}$, $where\ j = examinee\ (1, 2, 3, \dots, N)$

- **Lord-Wingersky formula** (dichotomous item model)

```
/* n: # of items
/* prob_vector : a vector of probabilities for an examinee

start f_rx(n ,prob_vector);
  f_rx_matrix_old = j(1+1,1,.);
  f_rx_matrix_old[1]= 1-prob_vector[1];
  f_rx_matrix_old[2]= prob_vector[1];

  do r = 2 to n;
    f_rx_matrix = j(r+1,1,.);
    f_rx_matrix[1]= f_rx_matrix_old[1]#(1-prob_vector[r]);
    f_rx_matrix[r+1]=f_rx_matrix_old[r]#prob_vector[r];
    do x = 2 to r;
      f_rx_matrix[x]= f_rx_matrix_old[x-1]#prob_vector[r]+f_rx_matrix_old[x]
                #(1-prob_vector[r]);
    end;

    f_rx_matrix_old = f_rx_matrix;
   end;
  return (f_rx_matrix_old);
finish;
```

Classification Decision Accuracy and Consistency in IRT by Using SAS/IML®

Seohee Park[1], Kyungyong Kim[1], and Sung-Hyuck Lee[2]

[1]University of Iowa and [2]ACT

Application

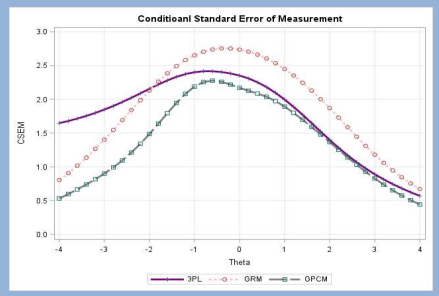
- Application of the program : Analysis of 2012 PISA reading test (booklet9).
- After using likewise deletion for missing data, 29,999 examinees' data were used.
- Booklet 9 consists of 8 passages (1 passage with 5 items, 4 passages with 4 items, and 3 passages with 3 items)
- Considering items in each passage as one polytomous item was used for GRM and GPCM models in order to reflect the violation of local independence, which is a fundamental assumption for 3PL model.
- Two categories were considered in this application (Pass/ Fail).
- The application used 17 different cut scores from -2 to 2 with 0.25 in order to investigate the effect from the location of cut score on CA/CC

Table 1. Cut scores for theta scale and total score scale

Theta scale		-2	-1.5	-1	-0.5	0	0.5	1	1.5	2
Total score scale	3PL	7	9	12	15	18	21	24	26	28
	GRM	5	7	10	13	17	20	23	25	27
	GPCM	3	6	9	13	17	21	24	26	28

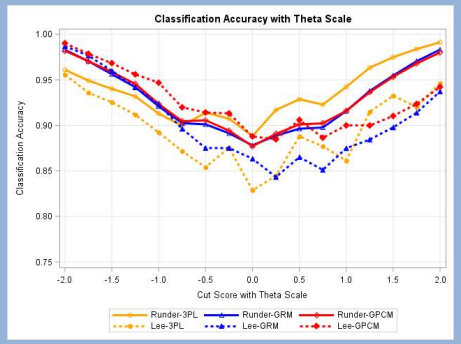
Note . Cut scores are ranged from -2 to 2 with 0.25 width. Only 9 cut scores are displayed because of the limit of space. Cut scores for the total score scale were converted from the theta scale by using the Lord-Wingkersky formula.

Result of Application



- Conditional Standard Error of Measurement (CSEM) is high in mid-values, and it is low in extreme values. CSEM with GPCM model is lower than other two methods.
- The point at which CSEM is high would have lower CA/CC indices.
- CA values are higher than CC regardless methods and models.
- Lines for three models with Runder's method less spread out than ones with Lee's method. It means Rudner method results in the similar CA/CC indices regardless models.

Figure 1. Standard Error of measurement for 41 points in theta scale.



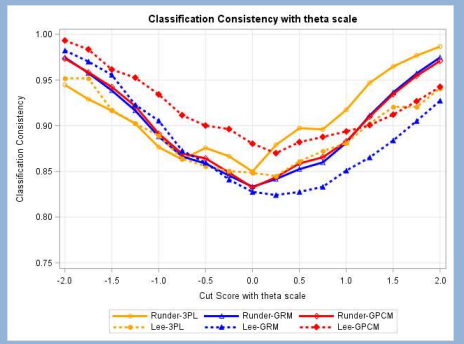


Figure 2. Classification Accuracy and Consistency

Note. Cut scores for Rudner method are in the theta scale, and cut scores for Lee method are in the total score scale. The converted cut scores for the total score scale in Table 1. were used.

REFERENCES

- Kolen, M. J., & Brennan, R. L. (2013). *Test equating: Methods and practices*. Springer Science & Business Media.
- Lathrop, Q. N., & Cheng, Y. (2013). Two approaches to estimation of classification accuracy rate under item response theory. *Applied Psychological Measurement*, 0146621612471888.
- Lee, W. C. (2010). Classification consistency and accuracy for complex assessments using item response theory. *Journal of Educational Measurement*, *47*(1), 1-17.
- Rudner, L. M. (2001). Computing the expected proportions of misclassified examinees. *Practical Assessment, Research & Evaluation*, *7*(4).
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement*.
- Wyse, A. E., & Hao, S. (2012). An evaluation of item response theory classification accuracy and consistency indices. *Applied Psychological Measurement*, 0146621612451522.

SAS® GLOBAL FORUM 2017

April 2 – 5 | Orlando, FL