A Model Evaluation When Associations Exists Across Testlets under Small Testlet Size Situations

Ou Zhang, University of Florida,
M. David Miller, PH.D., University of Florida,
Mac Cannady, Boston College

ABSTRACT

This study investigated the effectiveness of ability parameter recovery for two models to detect the influence of the association between testlets under the small testlet size situation. A simulation study was used to compare two Rasch type models, which were the Rasch tesetlet model and the Rasch subdimension model. The results revealed that the Rasch subdimension model performed better than the Rasch testlet model as the existence of between testlets association. The results also indicated that as the sample size increased, the discrepancies between model estimates and the real data set increased. The study concluded that using the Rasch subdimension model for testlet item analyses is efficient for small testlet size and non-adaptive typed tests when between testlets association exists. In sum, the Rasch subdimension model offered an advantage over the Rasch testlet model as it avoided standard error of measurement underestimation between testlets and better ability parameter estimations in the small testlet size situations.

Key Words: IRT, non-adaptive test, small testlet, model fit

INTRODUCTION

- A testlet, is a scoring unit, a set of items following the same prompt, within a test that is smaller than the whole test (Wainer & Kiely, 1987). Items within testlets are locally dependent because they are associated with the same stimulus.
- The National Board of Osteopathic of Medical Examiners (NBOME) offers computer-based COMLEX-USA exams online. The COMLEX-USA level-2 exam consists of 141 independent items and 209 testlet items grouped in 95 testlets. The testlet sizes range from 2 to 4 items per testlet (small testlet size). There are five item types throughout the test. Among all five-item types, there are 3 different types of testlet items (i.e. B, S, and F).
- Because some testlets may have similar item format (i.e. both belong to one of the testlet item types, like B, S, F) and they may share similar content subdomain.
- So, not only is there associations within each testlet, but also there are **possible associations** (denoted as *testlet correlation*) between two or more testlets.

THEORETICAL FRAMEWORK

Currently, the testlet model method is widely used for testlet analyses.

- In the Rasch testlet model, σ_θ^2 has to be set at unity for model identification (i.e. $\sigma_\theta^2 = 1$). One limit of the testlet model is that the model requires all the latent traits to be independent of one another. This constraint is too restrictive to allow for possible item association between testlets. Therefore, further exploration of the between testlets association is impossible in the testlet model.
- The subdimension model (Brandt, 2007a, 2008) has been proposed to solve the between testlets item association issue. The subdimension model is based on the assumption that each person has an overarching ability in the measured dimension (denoted as main dimension), and testlet effects (denoted as subdimensions) are independent of main dimension but allows for possible subdimension associations by constraining the sum of the testlet effects (i.e. subdimension effects) to zero.

RESEARCH DESIGN AND METHODS

Model Used to Generate Data for the Simulations

In order to quantify the extent of these variations local effect, **the Rasch subdimension model (Brandt, 2007a, 2008)** was appropriate for the data simulation.

Model's Main Dimension and Subdimension Covariance Matrix Definition

$$\Sigma = \begin{bmatrix} \sigma_\theta^2 & 0 & \cdots & 0 \\ 0 & \sigma_{\gamma_1}^2 & \cdots & \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma_{\gamma_D}^2 \end{bmatrix}$$

Rasch testlet model

$$\Sigma = \begin{bmatrix} \sigma_\theta^2 & 0 & 0 & 0 \\ 0 & \sigma_{\gamma_1}^2 & \sigma_{\gamma_1}\sigma_{\gamma_2} & 0 \\ 0 & \sigma_{\gamma_1}\sigma_{\gamma_2} & \sigma_{\gamma_2}^2 & 0 \\ 0 & -\sigma_{\gamma_1}\sigma_{\gamma_2} & -\sigma_{\gamma_1}\sigma_{\gamma_2} & \sigma_{\gamma_3}^2 \end{bmatrix}$$

Rasch subdimension model

Data Source and Population parameters

The population item parameters and ability parameters were randomly drawn from normal distributions for each condition .

$$\theta_j \sim (0,1), b_i \sim (0,1)$$

Parameter Estimation

The parameters of the dataset in 2 models were analyzed using Marginal Maximum Likelihood (MML) methods. The estimations of the simulees' abilities were calculated by Expected a Posteriori Estimation (EAP; Bock & Mislevy, 1982).

Statistical Software

The response data were generated using the statistical software **R 2.12.2**. The parameters of the dataset in 2 models were analyzed with **ConQuest Version 2.0**.

Simulation Design

Our study was a four-factor completely crossed design: 3 (testlet correlation changes) x 4 (levels of local dependence effect) x 3 (ratio of testlet items and independent items) x 2 (sample size).

1. The testlet sizes chosen were based on the sizes less often discussed in the applied literature. Thus, for the simplicity of the study, only one testlet size (testlet size: 5) was used.
2. Three different testlet correlations between similar testlet formats (i.e. B, S, F types) were applied (i.e. 0.1, 0.2, 0.3).
3. The ratio of the correlated/total testlet numbers is very important in research. However, for this simplicity of the study, only three correlated testlets were included in this study.
4. Four levels of local dependence effect were examined: (0.25,0.5,0.75,1.0).
5. Among all 60 items, the ratio of testlet items to independent items were 1:3, 1:1, 3:1.
6. Two different sample sizes of examinees (500,1000) were applied.

Analysis Criteria

The likelihood ratio test :

$$\chi^2(df_D) = -2[\ln L_{\mathrm{null}} - \ln L_A]$$

Akaike's information criterion (AIC):

$$AIC = -2\ln L + 2P$$

Bayesian Information Criterion (BIC):

$$BIC = -2\ln L + 2P\ln(N)$$

Bias:

$$bias_{\hat{\theta}} = \frac{\sum_{j=1}^{n}\hat{\theta}_j - \theta_j}{n}$$

Root Mean Square Error (RMSE) :

$$RMSE_{\hat{\theta}} = \sqrt{\frac{\sum_{j=1}^{n}(\hat{\theta}_j - \theta_j)^2}{n}}$$

Test Reliability:

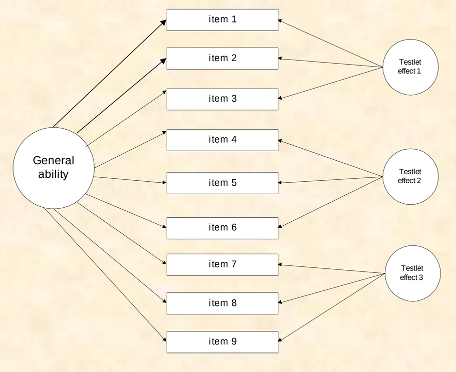
$$Test\ Reliability = \frac{Var(\theta_T)}{Var(\theta_{EAP})} = \frac{S^2(\hat{\theta}) - (\overline{s.e_\theta^2})}{S^2(\hat{\theta})}$$

IRT MODELS

Rasch Testlet Model

The Rasch testlet model includes a random effect parameter, which models the local dependence among items within the same testlet (e.g. Wang & Wilson, 2000). It can be written as

$$P_{ji1} = \frac{\exp(\theta_j - b_i + \gamma_{d(i)j})}{1+\exp(\theta_j - b_i + \gamma_{d(i)j})}$$



Rasch Testlet Model
(Wilson & Wang, 2000)

where P_{ji1} is the probability that examinee j answers item i correctly;

$\theta_j \sim N(0,1)$ is the ability of examinee j ;

$b_i \sim N(\mu_b, \sigma_b^2)$ is the difficulty of item i, and

$\gamma_{d(i)j} \sim N(0, \sigma_{\gamma_{d(i)}}^2)$ is a random effect that represents the interaction of person j with testlet $d(i)$ (i.e., testlet d that contains item i).

With j=1,…, J and J the total number of examinees,

Restriction 1: $\sigma(\theta_J, \gamma_{jd(i)}) = 0$ for all $d = 1,\ldots,D$ (1)

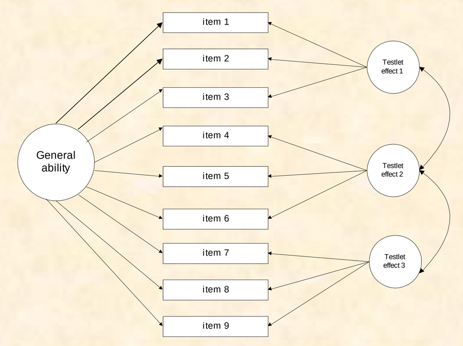
Restriction 2: $\sigma(\gamma_{jd(i)}, \gamma_{jd(i')}) = 0$ for all $d = 1,\ldots,D$ (2)

Restriction 3: $\sum_{j=1}^{J}\theta_j = 0$ (3)

Rasch Subdimension Model :

Brandt (2007a, 2008) proposed the Rasch subdimension model, which is similar to the Rasch testlet model (Wang & Wilson, 2005) in that it allows for association between testlet effects. It can be written as follows:

$$P_{ji1} = \frac{\exp(\theta_j - b_i + \gamma_{d(i)j})}{1+\exp(\theta_j - b_i + \gamma_{d(i)j})}$$



Rasch Subdimension Model
(Brandt, 2008)

where P_{ji1} is the probability that examinee j answers item i correctly;

$\theta_j \sim N(0,1)$ is the ability of examinee j ;

$b_i \sim N(\mu_b, \sigma_b^2)$ is the difficulty of item i, and

$\gamma_{d(i)j} \sim N(0, \sigma_{\gamma_{d(i)}}^2)$ is a random effect that represents the interaction of person j with testlet $d(i)$ (i.e., testlet d that contains item i). All the parameters in the model have the same definitions as the Rasch testlet model except Restriction 2.

Restriction 1: $\sigma(\theta_J, \gamma_{jd(i)}) = 0$ for all $d = 1,\ldots,D$ (4)

Restriction 2: $\sum_{d=1}^{D}\gamma_{jd(i)} = 0$ for all $j = 1,\ldots, J$. (5)

Restriction 3: $\sum_{j=1}^{J}\theta_{jd} = 0$ (6)

RESULTS (SELECTED) Rasch Testlet Model vs Rasch Subdimension Model-Deviance, AIC, BIC-Sample size 1000

	Rasch Testlet Model				Rasch Subdimension Model			
Condition	No.Parameters	mean.deviance	mean.AIC	mean.BIC	No.Parameters	mean.deviance	mean.AIC	mean.BIC
1	69	71091.4976	71229.4976	72044.7679	96	70944.8651	71136.8651	72271.1541
2	69	75077.7995	75215.7995	76031.0697	96	74944.4365	75136.4365	76270.7255
3	69	75249.1993	75387.1993	76202.4695	96	75063.5283	75255.5283	76389.8173
4	69	74986.5889	75124.5889	75939.8591	96	74833.9646	75025.9646	76160.2536
5	66	69053.3789	69185.3789	69965.2026	75	68910.8756	69060.8756	69947.0389
6	66	70932.7197	71064.7197	71844.5434	75	70871.5578	71021.5578	71907.7211
7	66	78140.6321	78272.6321	79052.4558	75	78011.5280	78161.5280	79047.6913
8	66	76443.1713	76575.1713	77354.9950	75	76362.0526	76512.0526	77398.2159
9	63	74116.0070	74242.0070	74986.3842	63	73992.1193	74118.1193	74862.4965
10	63	78042.0984	78168.0984	78912.4755	63	78001.2896	78127.2896	78871.6668
11	63	72976.9271	73102.9271	73847.3043	63	72910.8105	73036.8105	73781.1876
12	63	76790.1390	76916.1390	77660.5162	63	76696.8423	76822.8423	77567.2195
13	69	70074.4890	70212.4890	71027.7592	96	69914.2401	70106.2401	71240.5291
14	69	70249.3276	70387.3276	71202.5978	96	70047.3362	70239.3362	71373.6253
15	69	75235.1419	75373.1419	76188.4121	96	75098.1554	75290.1554	76424.4444
16	69	76495.1190	76633.1190	77448.3892	96	76342.3972	76534.3972	77668.6862
17	66	72001.4514	72133.4514	72913.2751	75	71867.5920	72017.5920	72903.7553
18	66	72541.5484	72673.5484	73453.3721	75	72421.9628	72571.9628	73458.1261
19	66	74068.5130	74200.5130	74980.3367	75	73965.6535	74115.6535	75001.8168
20	66	77324.8610	77456.8610	78236.6847	75	77160.9941	77310.9941	78197.1574
21	63	72166.4972	72292.4972	73036.8744	63	72042.1848	72168.1848	72912.5619
22	63	75975.0048	76101.0048	76845.3819	63	75858.9916	75984.9916	76729.3688
23	63	74450.6895	74576.6895	75321.0667	63	74376.2149	74502.2149	75246.5921
24	63	76122.1356	76248.1356	76992.5127	63	76075.1884	76201.1884	76945.5655
25	69	72099.6201	72237.6201	73052.8904	96	71960.5665	72152.5665	73286.8555
26	69	71091.7240	71229.7240	72044.9942	96	70882.6840	71074.6840	72208.9730
27	69	74892.3702	75030.3702	75845.6404	96	74742.6033	74934.6033	76068.8923
28	69	76072.7804	76210.7804	77026.0506	96	75860.4226	76052.4226	77186.7116
29	66	68629.0503	68761.0503	69540.8740	75	68492.2570	68642.2570	69528.4203
30	66	76801.3774	76933.3774	77713.2011	75	76674.5828	76824.5828	77710.7461
31	66	75584.4791	75716.4791	76496.3028	75	75449.2741	75599.2741	76485.4374
32	66	77454.3587	77586.3587	78366.1824	75	77361.6240	77511.6240	78397.7873
33	63	73076.3384	73202.3384	73946.7155	63	75309.5762	75435.5762	76179.9534
34	63	74346.0333	74472.0333	75216.4105	63	77949.3777	78075.3777	78819.7549
35	63	75960.0352	76086.0352	76830.4123	63	74515.0729	74641.0729	75385.4500
36	63	75983.1827	76109.1827	76853.5598	63	72604.7487	72730.7487	73475.1259

EMPIRICAL CASE

- The 2008 National Board of Osteopathic of Medical Examiners (NBOME) COMLEX-USA Level-2 exam data was used as an empirical case for this study. The item type was identified (i.e. A -single item, D-single Item with graph, B-matching item, S-testlet item, F-testlet item with graph). The B, S, and F type items were categorized as testlet items. A total of 450 examinees were included in the examinee population. No missing data existed. The data of the block-1 was used including 50 items categorized as 27 independent items and 23 testlet items within 10 testlets.
- The values of deviance for the Rasch testlet model and the Rasch subdimension model were 19,237.40 and 19,190.02, respectively.
- The values of AIC for these two models were 19357.40 and 19,310.02, respectively.
- The values of BIC of these two models were 19970.51 and 19923.13, respectively. (The total numbers of estimated parameters for these two models are 60 and 95.)
- The estimates of test reliability for the overarching latent trait are 0.891 for the Rasch testlet model, 0.882 for the Rasch subdimension model. Thus, the Rasch testlet model appeared to slightly overestimate the test reliability due to its ignorance of the association between testlets.

In summary, the Rasch subdimension model has a better fit, compared with the Rasch testlet model when used to analyze NBOME COMLEX exams.

DISCUSSION AND CONCLUSION

- The Rasch subdimension model performed better than the Rasch testlet model under small testlet sizes and when associations between testlets exist.
- Sample size had a observable effect on the analysis results for the Rasch subdimension model and the Rasch testlet model.
- No evident pattern can be found to reveal the association between the factor variations (i.e., the sample size, the association between testlets) and the bias/RMSE result.
- Although there was no obvious discrepancy of the test reliability estimates between the Rasch testlet model and the Rasch subdimension model, a small overestimation trend merged from the Rasch testlet model test reliability estimation.

REFERENCE (SELECTED)

- **Brandt, S. (2007a). Applications of a Rasch model with subdimensions. Paper presented at the 2007 Annual Conference of the American Educational Research Association (AERA), Chicago.**
- **Brandt, S. (2007b). Item bundles with items relating to different subtests and their influence on subtests' measurement characteristics. Paper presented at the 2007 Annual Conference of the American Educational Research Association (AERA), Chicago.**
- **Brandt, S. (2008). Modeling tests with subtests (Paper submitted for publication).**
- **Wainer, H. & Kiely, G, L. (1987). Item clusters and computerized adaptive testing: A case for testlets. Journal of Educational Measurement, 24, 185-201.**
- **Wainer. H., Lewis. C. (1990). Toward a Psychometrics for Testlets. Journal of Educational Measurement. 27(1), 1-14.**
- **Wang, W.-C., & Wilson, M. (2005). The Rasch testlet model. Applied Psychological Measurement, 29(2), 126–149.**
- **Zhang, O., Shen, L., Cannady, M. (2010). Polytomous IRT or Testlet Model: An Evaluation of Scoring Models under Small Testlet Size Situation".Paper presented at The 15th International Objective Measurement Workshop (IOMW 2010), Boulder.**

ACKNOWLEDGEMENT & CONTACT INFORMATION

- I would like to express my sincere appreciation to Dr. M. David. Miller, my advisor, for providing valuable guidance and continuous support. I would also like to thank Dr. James J. Algina, for sharing his ideas and corrections on this project.
- My deepest gratitude goes to Mac Cannady, for his help and assistance.
- Thanks to my summer internship mentor, Dr. Feiming Li (NBOME) and Vice President, Dr. Linjun Shen (NBOME) for giving me such a valuable opportunity to enter the educational measurement industry.

Ou Zhang
Research and Evaluation Methodology Program, College of Education
University of Florida
119G Norman Hall, Gainesville, FL 32611-7047
E-mail: zhango@ufl.edu