วารสารบัณฑิตศึกษา มหาวิทยาลัยราชภัฏวไลยอลงกรณ์ ในพระบรมราชูปถัมภ์
ปีที่ 11 ฉบับพิเศษ พฤศจิกายน 2560

EVALUATION OF ABILITY ESTIMATES UNDER BI-FACTOR TESTLET MODEL

Chalie Patarapichayatham and Akihito Kamata
Southern Methodist University, Dallas, Texas, USA

ABSTRACT

When a test is composed of testlets, such as items associated with common reading passages, a bi-factor testlet model has been proven to be effective in item-level test data analysis, as well as estimation of abilities. The objective of this study was to investigate the quality of ability estimates under some bi-factor testlet model specifications by focusing on the correlations between testlet factors though a simulation study and real data set. We focused on two models; (a) a conventional two-parameter bi-factor testlet model with zero correlations between factors, and (b) a modified two-parameter bi-factor model with freely estimated covariances between testlet factors. The results showed that the correlations between general factor and testlet factors were all positive. However, the correlations between testlet factors were all negative. Results from the analysis of the real data displayed similar patterns to the simulated data analysis.

บทคัดย่อ

โมเดลไบแฟคเตอร์เทสเลทมีความเหมาะสมในการวิเคราะห์ข้อมูลรวมถึงมีความเหมาะสมในการประมาณค่าเมื่อแบบทดสอบเป็นแบบเทสเลทตัวอย่างเช่นแบบทดสอบแบบบทความที่มีข้อคำถามที่เกี่ยวข้องภายใต้บทความนั้นๆ ความสามารถ การศึกษาในครั้งนี้มีวัตถุประสงค์เพื่อตรวจสอบคุณภาพการประมาณค่าความสามารถของโมเดลไบแฟคเตอร์เทสเลทโดยมุ่งเน้นที่ความสัมพันธ์ระหว่างเทสเลทแฟคเตอร์ผ่านการจำลองข้อมูลและข้อมูลจริง โมเดลหลักในการตรวจสอบคือ (1) โมเดลไบแฟคเตอร์เทสเลท 2 พารามิเตอร์แบบดั้งเดิมที่ความสัมพันธ์ระหว่างแฟคเตอร์เป็นศูนย์ (2) โมเดลไบแฟคเตอร์เทสเลท 2 พารามิเตอร์แบบประยุกต์ที่ความสัมพันธ์ระหว่างแฟคเตอร์ไม่เป็นศูนย์ ผลการศึกษาพบว่าค่าความสัมพันธ์ระหว่างแฟคเตอร์หลักและเทสเลทแฟคเตอร์เป็นบวกทั้งหมด ในขณะที่ค่าความสัมพันธ์ระหว่างเทสเลทแฟคเตอร์และเทสเลทแฟคเตอร์เป็นลบทั้งหมด นอกจากนี้ผลการวิเคราะห์ข้อมูลจริงสอดคล้องกับผลการวิเคราะห์ข้อมูลจำลอง

Keywords

bi-factor model, testlet model, bi-factor testlet model, testlet-based item response theory, reading comprehension

คำสำคัญ

โมเดลไบแฟคเตอร์ เทสเลทโมเดล โมเดลไบแฟคเตอร์เทสเลท

INTRODUCTION

A testlet-based item response theory (IRT) model can be a valuable modeling framework for educational assessment for several reasons. First, testlet IRT model may provide more information than conventional IRT models. For example, a researcher may be interested in the predictive relationships between testlet and general factors. Also, with a testlet model, we can directly examine the strength of associations between testlet factors and their associated items, which is simply done by evaluating factor loadings. Finally, testlet models can be easily applied to various educational measurement practices, such as computerize adaptive testing (CAT) and test equating. For these reasons, its utilities have been investigated and documented by many authors (e.g., Chen et al., 2006; DeMars, 2006; Li, Bolt & Fu, 2005; Li, Bolt & Fu, 2006). However, much attention has not been given to associations between testlet and general factors, probably because the bi-factor testlet model assumes testlet factors are nuisance factors and uncorrelated to the general factor. However, if testlet factors are in fact correlated with the general factor, it raises a question regarding their validity depending on the direction and magnitude of the correlation. This study numerically investigated the quality and characteristics of ability estimates under the bi-factor testlet IRT model by focusing on the correlations between testlet and general factors. More specifically, correlations between ability estimates for testlet and general factors were investigated. We also investigated another form of testlet scores, which was the sum of general factor and the testlet factor scores under the conventional two-parameter bi-factor IRT model and the modified two-parameter bi-factor IRT model. We also evaluated correlation between observed testlet total scores and total score for the entire test.

METHODS

Simulation Design

3 magnitudes of testlet effect [very small (0.0001), small (0.1) and large (0.4) in the standardized general factor ability scale], 2 magnitudes of discrimination power [weak (0.5) and strong (1.5) in the logistic scale], and 2 sample sizes [small (500), and large (1,000)], totaling 12 simulation conditions were considered in this study. The simulation factors were conducted based on our previous investigation. It was assumed that the test was consisted of 3 testlets, where each testlet contained 5 dichotomously

scored items with item difficulties ranging from -1.4 to 1.4. These item difficulties were fixed for all 12 simulation conditions. It was assumed that examinees' abilities were randomly sampled from the standard normal distribution. It was also assumed that the testlet effects were normally distributed with mean of zero and variance of either 0.0001, 0.1 or 0.4, depending on the simulation condition. Based on these specifications, dichotomous item response data were randomly generated by the 2PL conventional testlet model with zero covariances between all factors. We evaluated how the estimated factor scores were correlated under each of the conventional two-parameter bi-factor testlet model and the modified two-parameter bi-factor testlet model. We also evaluated correlations between subscale testlet raw score and testlet total scores. Data were generated by R and parameter estimations were conducted by M*plus* software.

Modeling

This study focuses on two bi-factor testlet IRT models. The first model is a conventional two-parameter bi-factor IRT model. This model can be viewed as a direct extension of the 2PL IRT model with additional secondary factors that represent testlet effects. A factor corresponding to ability in the 2PL IRT model is referred to as a general factor. On the other hand, a factor associated with each testlet is referred to as a specific factor or a testlet factor. In this model, a covariance between any two factors is constrained to be zero, and we refer a model with this constraint to as the conventional two-parameter bi-factor testlet IRT model. The second model is a modified version of the first model, where the covariances between testlet factors are freely estimated. We refer this model to as the modified two-parameter bi-factor testlet IRT model.

Real Data Study

A reading comprehension data for 1,000 5th grade students were analyzed. The data contained 6 testlets, in which 9 dichotomously scored items were associated within each testlet, totaling 54 items. Data were collected in spring 2013-14 school year.

Modeling

Same models were investigated for real data analysis: (a) a conventional two-parameter bi-factor IRT model, and (b) a modified two-parameter bi-factor testlet IRT model.

RESULTS

The correlations between subscale testlet raw scores were all positive, ranging 0.144 to 0.653. On the other hand, the correlations between testlet factor scores were all estimated to be negative across all simulation conditions. The magnitudes ranged from -0.451 to -0.110 for the conventional two-parameter bi-factor testlet IRT model, while they were from -0.993 to -0.021 for the modified two-parameter bi-factor testlet IRT model. It was evident that the correlations between testlet factor scores were different from the correlations between subscale testlet raw scores. Our results indicated that absolute values of the magnitudes from subscale testlet raw scores and conventional two-parameter bi-factor testlet IRT model were similar to each other. However, the absolute values of magnitudes from modified two-parameter bi-factor testlet IRT model were much higher. On the other hand, the correlations between subscale testlet raw scores and total testlet raw scores were in the range of 0.637 to 0.901, while the correlations between testlet factor scores and general factor scores were all positive, raging 0.006 to 0.611. It was evident that testlet factor scores correlated with general factor scores much lower than testlet raw scores did. It is also quite evident that testlet factor scores themselves are not appropriate to be used as subscale scores to represent performance on testlets.

Table 1 : Correlations between subscale testlet raw scores and correlations between subscale testlet raw scores and total testlet raw scores

Condition	T_1T_2	T_1T_3	T_2T_3	T_1T_T	T_2T_T	T_3T_T
1	0.201	0.230	0.207	0.688	0.682	0.698
2	0.271	0.232	0.229	0.696	0.722	0.695
3	0.653	0.505	0.595	0.842	0.898	0.809
4	0.646	0.528	0.640	0.835	0.901	0.837
5	0.186	0.212	0.269	0.678	0.707	0.696
6	0.246	0.294	0.227	0.729	0.689	0.712
7	0.591	0.483	0.583	0.813	0.882	0.817
8	0.574	0.489	0.620	0.806	0.883	0.832
9	0.180	0.161	0.239	0.637	0.710	0.691
10	0.203	0.144	0.181	0.667	0.701	0.646
11	0.438	0.454	0.425	0.786	0.801	0.786
12	0.449	0.389	0.441	0.768	0.818	0.770

Note. T_1T_2 = correlations between subscale testlet 1 raw scores and subscale testlet 2 raw scores, T_1T_3 = correlations between subscale testlet 1 raw scores and subscale testlet 3 raw scores, and T_2T_3 = correlations between subscale testlet 2 raw scores and subscale testlet 3 raw scores, T_1T_T = correlations between subscale testlet 1 raw scores and total testlet raw scores, T_2T_T = correlations between subscale testlet 2 raw scores and total testlet raw scores, and T_3T_T = correlations between subscale testlet 3 raw scores and total testlet raw scores.

Given the correlations between testlet raw scores as the base line, our results raised a question why the correlations between testlet factor scores and general factor scores were quite different from the correlations between testlet raw scores. Another question was whether we could use testlet factor scores or general factor scores as trait level estimates. It is speculated that either testlet factor scores or general factor scores may be affected by negative correlations between testlet factor scores. The correlations between subscale testlet raw scores were all positive, and the correlations between subscale testlet raw scores and total testlet raw scores were high. On the other hand, since the correlations between testlet factor scores were all negative, the correlations between testlet factor scores and general factor scores turned out to be much lower. It is evident that general factor scores may not be affected by negative testlet factor scores correlations. On the other hand, it is quite clear that testlet factor scores are affected from negative testlet factor scores correlations. For this reason, a question is; does it make sense to use testlet factor scores as the testlet subscale scores itself? Our results indicated that probably it does not make sense, simply because in reality the correlations between testlet factor scores should behave like correlations between testlet raw scores. Our concern is that having negative correlations between testlet factor scores may affect any interpretations based on testlet model. Also, if testlet factor scores are used as testlet subscale scores, it would raise a question regarding their validity.

We also investigated a linear combination of a general factor score and a testlet factor score as the testlet subscale score. Our investigation demonstrated that the correlations between the linear combination and general factor scores were all positive in the range of 0.844 to 0.998. These results were reasonable than the correlations between testlet factor scores themselves and general factor scores. This may be an indication that a use of a linear combination of the general factor scores and testlet factor scores as the testlet subscale scores may make more sense. However, correlations between the linear combination and the general factor scores were slightly higher than we expected. Our anticipation was that they should appear close to the correlations between subscale testlet raw scores and total testlet raw scores.

Table 2 : Correlations between a linear combination of general factor scores and testlet factor scores and general factor scores

	Conventional bi-factor model			Modified bi-factor model		
Condition	$(G_\theta+T_{\theta1})*G_\theta$	$(G_\theta+T_{\theta2})*G_\theta$	$(G_\theta+T_{\theta3})*G_\theta$	$(G_\theta+T_{\theta1})*G_\theta$	$(G_\theta+T_{\theta2})*G_\theta$	$(G_\theta+T_{\theta3})*G_\theta$
1	0.998	0.996	0.999	0.995	0.992	0.998
2	0.997	0.998	0.970	0.991	0.993	0.984
3	0.998	0.997	0.993	0.995	0.990	0.993
4	0.991	0.996	0.992	0.993	0.991	0.994
5	0.857	0.994	0.997	0.931	0.975	0.987
6	0.998	0.996	0.998	0.997	0.994	0.997
7	0.990	0.997	0.993	0.991	0.995	0.993
8	0.982	0.991	0.996	0.989	0.997	0.997
9	0.992	0.998	0.996	0.998	0.996	0.994
10	0.910	0.986	0.885	0.933	0.960	0.906
11	0.981	0.878	0.968	0.961	0.913	0.958
12	0.945	0.936	0.934	0.844	0.976	0.919

Note. $(G_\theta+T_{\theta1})*G_\theta$= correlations between a linear combination of general factor scores and testlet 1 factor scores and general factor scores, $(G_\theta+T_{\theta2})*G_\theta$= correlations between a linear combination of general factor scores and testlet 2 factor scores and general factor scores, and $(G_\theta+T_{\theta3})*G_\theta$= correlations between a linear combination of general factor scores and testlet 3 factor scores and general factor scores.

Results from real data set was as followed. The correlations between subscale testlet raw scores were all positive in the range of 0.434 to 0.654. On the other hand, the correlations between testlet factor scores were all negative ranging from -0.313 to -0.069 by the conventional two-parameter bi-factor testlet model, while they were -0.826 to -0.076 by the modified two-parameter bi-factor testlet model. The correlations between subscale testlet raw scores and total testlet raw scores were in the range of 0.757 to 0.850, whereas the correlations between testlet factor scores and general factor scores were in the range of 0.003 to 0.159. Also, the correlations between the linear combination and general factor scores were all positive in the range of 0.921 to 0.989. Overall, it was confirmed that results from the analysis of the real data displayed similar patterns to the simulated data analysis.

Regarding effect of simulation factors, our investigation demonstrated that the magnitudes of the correlations between subscale testlet raw scores, as well as the correlations between subscale testlet raw scores and total testlet raw scores, became larger when the item discrimination was larger. It is not surprising because the 2PL ability estimates are weighted by item discrimination. On the other hand, the correlations between a linear combination of the general factor scores and testlet factor scores and the general factor scores became smaller when testlet effect was larger. Since testlet effect affects the probability of each individual answers each item correctly, when testlet effect is small, the variation between individuals would be very small. In this case, the correlations between the linear combination and the general factor scores would likely behave similar to the correlations between general factor scores, which should be very high. On the other hand, when testlet effect is larger, the variation between individuals would be larger. Therefore, a linear combination of general factor scores and testlet factor scores was likely affected by the testlet effect.

CONCLUSIONS AND NEXT STEPS

This paper investigated how testlet factor scores and ability estimates were correlated each other under two bi-factor testlet IRT model specifications though a simulation study and real data from a reading comprehension test. First, the results showed that the correlation between testlet factor scores from two testlets were uniformly negative. This was contrary to positive correlation between testlet-level observed total scores. Second, it was found that the correlations between testlet factor scores and ability estimates were positive. However, the magnitude of the correlations were uniformly smaller than the correlations between testlet-level observed total scores and the test-level observed total scores. Lastly, the sum of ability score and testlet factor score correlated highly with the ability score. This may be an indication that a use of a linear combination of the general factor scores and testlet factor scores as the testlet subscale scores may more make sense. Therefore, it is our intention to further investigate this issue. The further investigation is warranted.

REFERENCES

Chen, F. F., West, S. G., & Sousa K. H. (2006). A Comparison of bifactor and second-order models of quality of life. **Multivariate Behavioral Research**, 41(2), 189–225.

DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. **Journal of Educational Measurement**, 43(2), 145-168.

Li, Y., Bolt, D. M., & Fu, J. (2005). A test characteristic curve linking method for the testlet model. **Applied Psychological Measurement**, 29(5), 340-356.

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. **Applied Psychological Measurement**, 30(3), 3-21.