Examining testlet effects on the PIRLS 2006 assessment

Yuwen Chang, National Taipei University of Education, yuwenc@ntue.edu.tw

Jennwu Wang, Fo-Guang University, cwwang@mail.fgu.edu.tw

Abstract

We fit 10 PIRLS testlets with standard item response and testlet response theory model. The variances of testlet effects are examined and compared. Item parameters and ability parameters estimated from the two different models are compared. The results indicate that variances of testlet effects range from .168 to .489. Item parameters as well as ability parameter estimated from two approaches correlated highly. When local dependence is ignored errors of estimation in the difficulty parameter and cutoff parameter are negligible. However, estimates of both discriminate parameter and guessing parameter are biased. The standard error of ability parameter is underestimated when the local independence is incorrectly assumed.

Keywords: *testlet response theory, graded response model, PIRLS, SCORIGHT*

Local independence is one of important assumption in standard item response theory. This assumption is not true for most of reading comprehension tests. In a reading comprehension test, a small number of passages followed several items that are paired with each passage. A group of items based on a common stimulus (passage) is called a testlet (Wainer & Kiely, 1987). Responses to items within a testlet are not likely to be independent of one another due to factors associated with the passage, such as subject matter expertise, misinterpretation of the passage, and so on.

It is well established that ignoring the testlet effect may result in underestimated standard error of ability parameter and biased item parameters (Ip, 2000; Sireci, Thissen, & Wainer, 1991; Wainer, 1995; Wainer, Bradlow, & Du, 2000; Wainer & Lukhele, 1997; Wainer & Thissen, 1996; Wainer & Wang, 2000; Wilson & Adams, 1995; Yen, 1993). Thus, several methods have been proposed for modeling these dependencies. Testlet response theory is one of such approach. Bradlow, Wainer, and Wang（1999）modeled the dependency because of testlet by adding a parameter to two-parameter item response theory. Wainer, Bradlow, and Du (2000) proposed a three-parameter logistic model for testlet-based test. Wang, Bradlow, and Wainer (2002) extended the previous models to include two basic probability kernels: the three-parameter logistic models for binary data and the graded response model introduced by Samejima (1969). They called it a general Bayesian model for testlets.

The purposes of the study are: (1) to examine the size of testlet effects in PIRLS assessment and explore if the size depends on types of passage; (2) to study consistency and standard errors of item and ability parameters estimated from the testlet response theory and standard item response model.

Method

Data sources

The study was based on the Progress in International Reading Literacy Study (PIRLS) database compiled by the International Association for the Evaluation of Educational Achievement that consists of nationally representative responses by fourth-grade students collected in 45 participants in the year 2006 (Mullis, Kennedy, Martin, & Sainsbury, 2004). The PIRLS database was collected to provide internationally comparative data about students' reading achievement in primary school. Two reading purposes (literary and informational) and a range of four comprehension processes are the foundation of the PIRLS

2006 assessment. The assessment consisted of ten passages, five for the literary purpose (labeled L1-L5), and five for the informational purpose (labeled I1-I5). Each passage was accompanied by approximately 12 test items, with about half in the multiple-choice format and half in the constructed-response format. The passages and items were distributed across 13 test booklets (see table 1). Each student only responded to one booklet. To ensure that there are not too many missing data for each student, only students who responded to one of the following booklets were included: booklets 1, 2, 3, 5, 6, 7 and Reader. Subjects who took first, second, and third booklets are combined into literary dataset (L dataset: L1-L4) and subjects who took fifth, sixth, and seventh booklets are combined into informational data set (I dataset: I1-I4). The third data set is Reader booklet (R dataset: L5, I5). The sample sizes for L, I and R datasets are 913, 923, and 914 respectively. The item numbers and formats are presented in table 2.

[Take in Table 1 about here]

[Take in Table 2 about here]

Data analysis

Since two question formats are used in the PIRLS 2006 assessment- multiple-choice and constructed-response, model that can deal with mixture item format is required. Bayesian model for testlets (Wang, Bradlow, & Wainer, 2002) thus is used for the study. In the general Bayesian model for testlets, the probit t_{ij} is forrmulated as

$$t_{ij} = a_j(\theta_i - b_j - \gamma_{id(j)})$$

Where a_j, b_j, and θ_i are item slope, item difficulty, and examinee proficiency, and $\gamma_{id(j)}$ is the testlet effect of item j to person i. When the testlet parameter is set to zero, all items are assumed to be independent.

Computer program Scoright 3.0 (Wang, Bradlow, & Wainer, 2005) was used to estimate item, ability and testlet parameters in two conditions- with and without testlet parameters in the model. In the program, the general Bayesian model for testlets is estimated using Markov Chain Monte Carlo methods. Details of the techniques are presented in Wang et al. (2002). The ability θ, $\gamma_{id(j)}$, and the item difficulty parameters were given the following normal priors:

$$a_j \sim N(\mu_a\ ,\ \sigma_a^2),\ b_j \sim N(\mu_b\ ,\ \sigma_b^2),\ q_j \sim N(\mu_q\ ,\ \sigma_q^2)\ ,\theta_i \sim N(0\ ,\ 1),\gamma_{id(j)} \sim N(0\ ,\ \sigma_{d(j)}^2)$$

All hyperparameters were given noninformative priors:

$$\mu_a \sim N(0\ , V_a),\ \mu_b \sim N(0\ , V_b),\ \mu_q \sim N(0\ , V_q),\ \text{and}\ |V_a|^{-1}=|V_b|^{-1}=|V_q|^{-1}=0$$

$\sigma_a^2 \sim \chi^{-2}_{g_a}$, $g_a = \frac{1}{2}$, $\sigma_b^2 \sim \chi^{-2}_{g_b}$, $g_b = \frac{1}{2}$, $\sigma_q^2 \sim \chi^{-2}_{g_q}$, $g_q = \frac{1}{2}$,

The distribution for $\sigma_{d(j)}^2$ is

$$\sigma^2_{d(j)} \sim \chi^{-2}_{g\tau}, \quad g\tau = \frac{1}{2}$$

Since the scale of the model is fixed by the unit variance of the ability distribution, the size of testlet variances can be interpreted by comparing to 1. A testlet effect of 1 means that the testlet variance is of the same magnitude as the variance of examinees.

For each condition, three chains stated at random were run. All chains had the same number of burin-ins (i.e. 7000). The number of iterations is 10000. The draws after 7000 were used for inference. The number of testlets for the L and for the I dataset is 4, while that for R dataset is 2. In the standard IRT condition, number of testlets for all datasets is set to zero.

The correlations and standard errors of item and ability parameters estimated from the two different models are calculated and compared. In addition, testlet effects for passages in each dataset are estimated and compared.

Finding and Discussion

The size of testlet effects was greater in the informational testlets than in the literary testlets as shown in the Table 3. In addition, in each kind of testlet, some passages have much higher conditional dependence than others. The variance of the testlet effect for I_1 (the content of the passage is 'Antarctica ') is .489, which is much greater than for other passages. We have no explanation for this at this moment. How does local dependence affect the estimation of item and ability parameters?

Since the effects of local dependence on item estimation are similar in the literary, informational, and Reader testlets. We present the results together. In Figures 1 and 5 are the estimated discriminations for two kinds of testlets using the standard IRT and testlet model? The results show that the item discriminations are somewhat overestimated by the standard IRT model. The findings are consistent with previous study (Wainer & Wang, 2000). The correlation between item parameters estimated from the two models is shown in Table 4. The correlation for discrimination parameters is around .98-.99. Similar pattern was found in the reader dataset.

Figure 2 and Figure 6 show the estimated values of difficulty parameter for literary and information items respectively. The estimates for both kinds of testlets are essentially identical for the two procedures. It appears that local dependence has negligible effect on the estimation of item difficulty parameters. The correlation for difficulty parameters is .99.

Again, the reader dataset presents similar results.

Figure 3 and figure 7 show the estimated values of the guessing parameter for literary and informational testlets. The value of the guessing parameter is slightly overestimated for literary items and underestimated for informational items. The correlation for guessing parameters is more variable in the literary testlets than in the informational testlets. It was found that the MCMC procedure does not estimate guessing parameters very well (Wang, Bradlow, & Wainer, 2002). The correlation for guessing parameters in the reader dataset is .825. The lower value of correlation may be due to small number of item.

Figures 4 and 8 are the estimated cutoff parameters. As with difficulty parameters, the estimates almost are identically with the two models.

[Take in Table 3 about here]

[Take in Figure 1-8 about here]

It has been shown that there are small but reliable biases in the estimated values of the item parameters of PIRLS testlet items when local dependence is ignored. How is the accuracy of the estimate of the parameter θ affected by the violation of local independence? Figure 9 and figure 10 show the comparisons of the standard errors of the θs estimated from the two models. Consisted with previous findings, the precision of the parameter θ will be overestimated when the testlet effect is ignored.

[Take in Figure 9 & 10 about here]

Conclusion and Implications

The examination of PIRLS testlets has provided some interesting results. We have found that the range of variation of the variance of the testlet effect for the different passages is around 0.3. Conditional dependence seems to have almost no effect on the estimation of item difficulty and of cutoff parameter. However, it tends to result in a bias on the estimation of *a* and *c* parameter. In addition, the precision of examinee proficiency is overestimated when conditional independence is incorrectly assumed.

Our analysis is based on PIRLS 2006 data from Taiwan. It would be interesting to know whether the findings will be consistent among the data sets from different countries. The nature of test items and the characteristics of passages may affect the amount of within-testlet dependency. What causes dependence is an important question needed to be further studied.

References

Bradlow, E. T., Wainer, H. & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64(2),* 153-168.

Campbell, JR., Kelly, D.L., Mullis, I. V. S., Martin, M. O., & Sainsbury, M. (2001). *Framework and specifications for PIRLS assessment 2001 (2nd ed.).* Chestnut Hill, MA: Boston College.

Dresher, A. R. (2003). The examination of local item dependency of NAEP assessments using the testlet model. Unpublished doctoral dissertation, University of Pittsburgh, Pittsburgh, Pennsylvania.

Foy & Kennedy (2008). PIRLS 2006 User Guide for the International Database. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Ip, E. H. S. (2000). Adjusting for information inflation due to local dependency in moderately large item clusters. *Psychometrika, 65,* 73-91.

Mullis, I. V. S., Kennedy, A. M.., Martin, M. O., & Sainsbury, M. (2004). *PIRLS 2006 Assessment Framework and Specifications.* TIMSS & PIRLS International Study Center, MA: Boston College.

Mullis, I. V. S., Martin, M. O., Kennedy, A. M.., & Flaherty, C. L. (Eds.) (2002). *PIRLS 2001 encyclopedia: A reference guide to reading education in the countries participating in IEA's Progress in International Reading Literacy Study (PIRLS).* Chestnut Hill, MA: Boston College.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16(2),* 159-176.

Muraki, E., & Bock, R. D. (2003) *PARSCALE: Parameter scaling of rating data.* Chicago, IL: Scientific Software.

Samejima, F. (1969). Estimation of latent ability usinig a response pattern of graded scores. Psychometrika Monographs, No. 17.

Sireci, S. G.., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28,* 237-247.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 Law School Admissions Test as an example. *Applied Measurement in Education, 8,* 157-186.

Wainer, H., & Lukhele, R. (1997). How reliable are TOEFL scores? *Educational and Psychological Measurement, 57,* 749-766.

Wainer, H., & Thissen, (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice, 15(1),* 22-29.

Wainer, H.& Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement, 37*,203-220.

Wainer, H., Bradlow, E. T., & Du. Z. (2000). Testlet response theory. An analog for the 3PL useful in testlet-based adaptive testing. In W. J. van der Linden & C. A. S. Glas (Eds.). *Computerized adaptive testing: Theory and practice (pp.245-270).* Boston, MA: Kluwer-Nijhoff.

Wainer, H., Sireci, S., & Thissen, D. (1991). Differential testlet functioning: Definition and detection. *Journal of Educational Measurement, 28*,197-219.

Wang, X., Bradlow, E. T., & Wainer H. (2002). A general Bayesian model for testlets: Theory and application. *Applied Psychological Measurement,26*, 109-128.

Wang, X., Bradlow, E. T., & Wainer H. (2005). User's Guide for SCORIGHT (version 3.0): A computer program for scoring tests built of testlets including a module for covariate analysis. Research report No. RR-04-49. Princeton, NJ: Educational Testing Service

Wang, W.C. & Wilson, M. R. (2005). Rasch testlet model. *Applied Psychological Measurement, 29(2)*, 126-149.

Yen, W. (1993). Scaling performance assessment: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187-213.

Table 1 PIRLS 2006 student booklet design

Booklet	1	2	3	4	5	6	7	8	9	10	11	12	R
Literary	L1	L2	L3	L4	I1	I2	I3	I4	L1	I2	L3	I4	L5
Informative	L2	L3	L4	I1	I2	I3	I4	L1	I1	L2	I3	L4	I5

Table 2 Item numbers and format for 10 passages

	L_1	L_2	L_3	L_4	I_1	I_2	I_3	I_4	L_5	I_5
Multiple-Choice	6	7	8	7	4	6	7	6	6	6
Constrcuted Dichotomous	4	3	2	1	3	2	2	2	3	6
Constrcuted Polytomous	3	3	4	4	4	4	3	4	3	2
Total	13	13	14	12	11	12	12	12	12	14

Table 3 The size of testlet effects for datasets

	L_1	L_2	L_3	L_4	I_1	I_2	I_3	I_4	R_L	R_I
Testlet Effects Var(γ)	.260	.168	.180	.245	.489	.238	.231	.322	.208	.192

Table 4 Correlation between parameters obtained from standard IRT and testlest model

	a	b	c	g_r		a	b	c	g_r
L_1	.977	.997	.773	.996	I_1	.993	.998	.995	.997
L_2	.997	.999	.966	1.0	I_2	.995	1.0	.984	.999
L_3	.979	.999	.966	.999	I_3	.989	.999	.997	.999
L_4	.988	.999	.957	.996	I_4	.982	.998	.944	.991
Literary Total	.981	.999	.912	.995	Information Total	.990	.999	.979	.997
L_5	.991	.997	.798	.989	I_5	.991	.999	.897	1.0
reader	.990	.999	.825	.999					

Figures 1-8

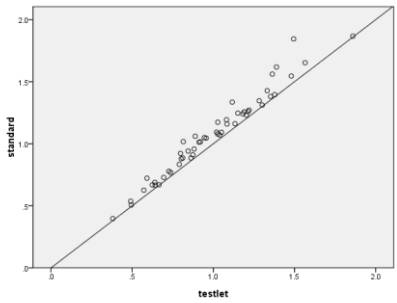

Figure 1 Estimated discrimination parameters for literary testlets obtained from standared IRT and testlet model

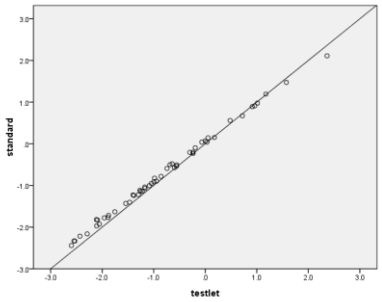

Figure 2 Estimated difficulty parameters for literary testlets obtained from standared IRT and testlet model

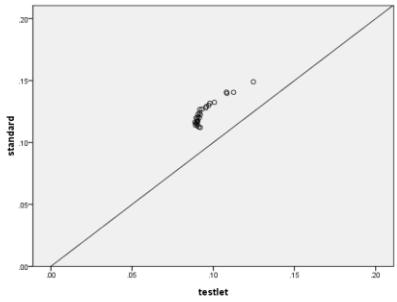

Figure 3 Estimated guessing parameters for literary testlets obtained from standared IRT and testlet model

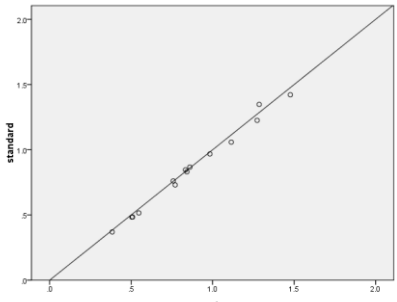

Figure 4 A comparison of the estimated vaules of cutoff parameter for Literary items

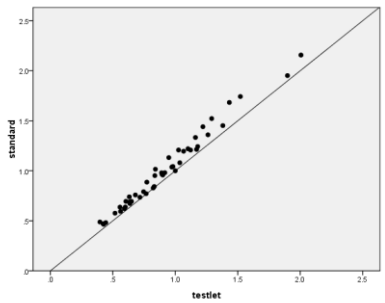

Figure 5 Estimated discrimination parameters for informational testlets obtained from standared IRT and testlet model

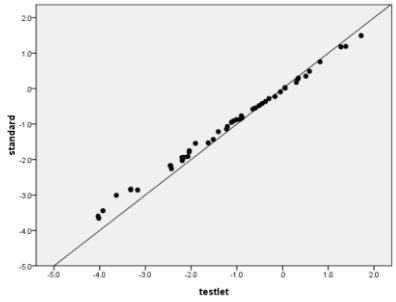

Figure 6 Estimated difficulty parameters for informational testlets obtained from standared IRT and testlet model

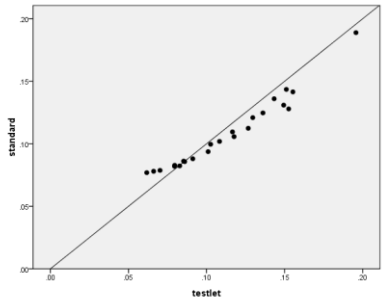

Figure 7 Estimated guessing parameters for informational testlets obtained from standared IRT and testlet model

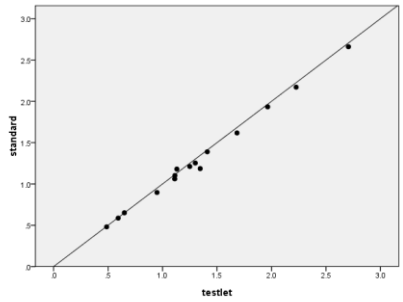

Figure 8 Estimated cutoff parameters for informational testlets obtained from standared IRT and testlet model

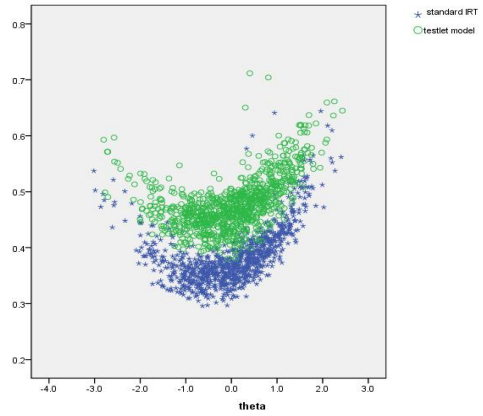

Figure 9 A Comparison of Standard Errors of Estimated Theta for Literary Passages

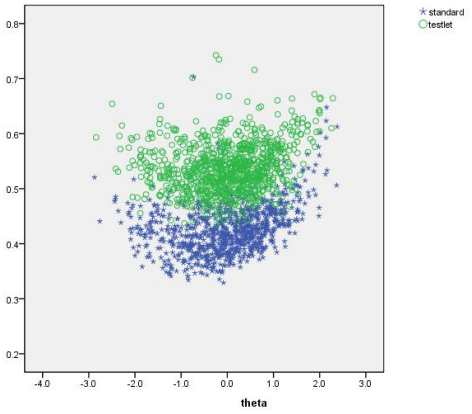

Figure 10 A comparison of Standard Errors of Estimated Theta for informational passages