Research Report

Application of a General Polytomous Testlet Model to the Reading Section of a Large-Scale English Language Assessment

Yanmei Li

Shuhong Li

Lin Wang

September 2010

ETS RR-10-21

Application of a General Polytomous Testlet Model to the Reading Section of a Large-Scale English Language Assessment

Yanmei Li, Shuhong Li, and Lin Wang
ETS, Princeton, New Jersey

September 2010

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Technical Review Editor: Daniel Eignor

Technical Reviewers: Weiling Deng and Insu Paek

Abstract

Many standardized educational tests include groups of items based on a common stimulus, known as *testlets*. Standard unidimensional item response theory (IRT) models are commonly used to model examinees' responses to testlet items. However, it is known that local dependence among testlet items can lead to biased item parameter estimates when using standard IRT models, and to overestimated reliability. In this study, a general polytomous testlet model was proposed to account for local dependence in testlet-based tests that contain both dichotomously and polytomously scored items. The proposed model and a standard IRT model were fit to simulated data and several real data sets from the reading sections of a large-scale English-language test, and model fit was evaluated. Item parameters and test information obtained from the two models were compared to check the impact of local item dependence. In addition, a multidimensional IRT model with simple structure was also fit to the real data sets. Results based on both simulated and real data suggested that local dependence had a small impact on item parameter estimates and a relatively larger impact on test information and reliability. It was also found that the multidimensional IRT model with simple structure fit the real data sets better than the general polytomous testlet model and the standard IRT model did.

Key words: item response theory, local dependence, polytomously scored items, reliability

Acknowledgments

The authors thank Shelby Haberman, Frederic Robin, and Alina von Davier for their valuable advice, and Cindy Nguyen for helping with the simulation data analyses. The authors also thank Dan Eignor, Weiling Deng, and Insu Paek for their helpful comments and suggestions.

Table of Contents

List of Figures

Page

List of Tables

Page

Standardized educational tests often include groups of items based on a common stimulus, known as *testlets.* When standard item response theory (IRT) models (such as the unidimensional two- or three-parameter models) are used to model examinees' responses to testlet items, one issue that is likely to arise is the violation of the local independence assumption of the IRT models. Local independence means that once the abilities influencing item performance are taken into account, examinees' responses to different items are statistically independent. When local independence does not hold, item responses are said to be locally dependent. For example, if one item contains relevant information for answering another item when several items refer to a common stimulus, as in the case of a testlet, local independence is violated. Previous studies have reported that applying a standard IRT model to testlet items while ignoring such local dependence could lead to biased item parameter estimates (Wainer & Wang, 2000) and to overestimated reliability (Sireci, Thissen, & Wainer, 1991). One approach to addressing this issue is to add a *random effect parameter* (also called a *testlet factor*) to standard IRT models so that local dependence among testlet items is taken into account (e.g., Bradlow, Wainer & Wang, 1999; Wainer, Bradlow, & Du, 2000; Wang, Bradlow, & Wainer, 2002). However, one limitation of the Bradlow et al. model is that it applies a common item discrimination parameter to both the general ability and testlet factors. Li, Bolt, and Fu (2006) investigated several alternative ways of accounting for local dependence in testlet-based tests and found that a general testlet model in which separate discrimination parameters were applied to the general ability and to the testlet factors provided a better fit to testlet data. The general testlet model is essentially the same as the *bifactor model* by Gibbons and Hedeker (1992). This model not only takes into account local dependence within the testlets but also provides more information about how items within a testlet are influenced by the testlet factor. Thus it appears to be a promising model for accounting for local dependence and for studying testlet effects.

The general testlet model considered in Li et al. (2006) only applies to dichotomously scored testlet items. However, many testlet-based tests contain both dichotomous and polytomous items. Therefore, it is necessary to extend the model to accommodate a mixed format test that contains a mixture of dichotomous and polytomous items. In this study, we examined the application of a general polytomous testlet model to the reading sections of a large-scale English language assessment in which items are typically presented in sets that are

associated with common stimuli (e.g., reading passages) and some of the items are polytomously scored (e.g., 0, 1, 2). It is important to evaluate whether local item dependence is present in such tests and whether it has any impact on various statistical results. Specifically, the general polytomous testlet model and a standard IRT model were fit to the data, and the model fit was evaluated. Item parameter estimates and test information obtained from the two models were compared to determine if any differences existed that would indicate possible impact due to local dependence within the testlets.

The General Polytomous Testlet Model

A general two-parameter normal ogive testlet model has been shown to be the best model for testlet-based tests among several alternative testlet models (Li et al., 2006). This model is formulated as

$$P(y_{ij}=1)=\Phi(a_{i1}\theta_j - t_i + a_{i2}\gamma_{d(i)j}), \qquad (1)$$

where $P(y_{ij}=1)$ is the probability that examinee j answers item i correctly; Φ denotes the cumulative distribution function (CDF) of a standard normal distribution; θ_j is the ability of examinee j; $\gamma_{d(i)j}$ represents a secondary dimension associated with testlet d (containing item i) for examinee j; t_i is a threshold parameter related to the difficulty of the item; and a_{i1} and a_{i2} indicate the discriminating power of an item with respect to θ and γ_d, respectively. The mean and variance of the distributions for both θ_j and $\gamma_{d(i)j}$ are fixed to $N(0,1)$ for identification purposes, and θ_j and $\gamma_{d(i)j}$ are assumed uncorrelated.

This model was extended to accommodate polytomous items, since many testlet-based tests contain both dichotomous and polytomous items. A generalized partial credit model (GPCM; Muraki, 1992) was utilized here. The extended testlet model can be expressed as

$$P_{ijk}=\frac{\exp\left[\Sigma_{v=0}^{k}\left(a_{i1}\theta_j - t_{iv} + a_{i2}\gamma_{d(i)j}\right)\right]}{\Sigma_{c=0}^{m_i}\exp\left[\Sigma_{v=0}^{c}\left(a_{i1}\theta_j - t_{iv} + a_{i2}\gamma_{d(i)j}\right)\right]}, \qquad (2)$$

where P_{ijk} is the probability of scoring in category k of the m_i+1 score categories of item i by examinee j, t_{iv} is the difficulty parameter for score category v of item i, and θ_j, $\gamma_{d(i)j}$, a_{i1}, and a_{i2} have the same interpretations as those in Equation 1. For notational convenience,

$\Sigma_{v=0}^{0}\left(a_{i1}\theta_j - t_{iv} + a_{i2}\gamma_{d(i)j}\right) \equiv 0$. Again, the mean and variance of the distributions for both θ_j and $\gamma_{d(i)j}$ are fixed to $N(0,1)$ for identification purposes, and θ_j and $\gamma_{d(i)j}$ are assumed to be uncorrelated.

The general polytomous testlet model was estimated using the SAS NLMIXED procedure (SAS Institute, 1999). The SAS NLMIXED procedure can be used to fit nonlinear mixed models, that is, models in which both fixed and random effects are permitted to have a nonlinear relationship to the response variable. Recent studies have shown that many common IRT models can be calibrated using the new NLMIXED procedure (e. g., De Boeck & Wilson, 2004; Rijmen, Tuerlinckx, De Boeck, & Kuppens, 2003; Sheu, Chen, Su, & Wang, 2005). The PROC NLMIXED fits nonlinear mixed models by maximizing an approximation to the likelihood integrated over the random effects (marginal maximum likelihood estimation). Two principal approximations to the integral are adaptive Gaussian quadrature and a first-order Taylor series approximation. A number of alternative optimization techniques are available to carry out the maximization; the default is a dual quasi-Newton algorithm. The Gauss-Hermite quadrature and dual quasi-Newton algorithm were used in this study. One attractive feature of the NLMIXED procedure is the easy implementation of a variety of models. However, for models with several random effects, the computational time is rather long. Appendix A gives a sample SAS code for fitting the general polytomous testlet model and the 2PL/GPCM.

Simulation Study

A simulation study was conducted to evaluate the parameter recovery of the general polytomous testlet model and a standard generalized partial credit model estimated using the NLMIXED procedure. The simulated test structure mimicked that of the real English language test analyzed in this study. Specifically, each simulated test contained 3 passages, with 14 items per passage. The last item in each passage was polytomously scored (0, 1, 2), while the remaining items were dichotomously scored (0, 1). Data were generated according to the general polytomous testlet model. Two levels of testlet effect (as measured by the item discrimination parameters with respect to the testlet factor) were studied: small and larger. The values in the simulation mimicked the testlet effect typically found in real data sets (e.g., Li et al., 2006; Wang et al, 2002). In the small testlet effect condition (Condition 1), the a_{i1} values were generated from a lognormal distribution such that the mean and standard deviation of a_{i1}

were 0.9 and 0.3, respectively, and the a_{i2} values were generated from a lognormal distribution such that the mean and standard deviation of a_{i2} were 0.4 and 0.2, respectively. In the larger testlet effect condition (Condition 2), a_{i1} values were generated from the same distribution as in Condition 1, while a_{i2} values were generated from a lognormal distribution such that the mean and standard deviation of a_{i2} were 0.8 and 0.2, respectively. For both conditions, the remaining parameters were generated as follows: $t_{iv} \sim N(0,1)$, $\theta_j \sim N(0,1)$, and $\gamma_{dj} \sim N(0,1)$. For each data set, 2000 examinees' item responses to the test were simulated. Ten data sets were simulated under each condition, resulting in a total of 20 simulated data sets.

The general polytomous testlet model was then fit to each data set. In addition, a 2PL and GPCM combination (2PL/GPCM) was fit to the data so that the differences between model parameters obtained from the true and alternative models can also be studied. In this study, we chose the 2PL model instead of the 3PL model for the dichotomous items based on previous research on model fit and computation issues related to the *c parameters* in the 3PL model. It is known that the pseudo-guessing parameters, c, in the 3PL model may not be well estimated because of a lack of information at the low end of the ability scale (Lord, 1980). The poorly estimated c parameters may further affect the estimation of other item parameters and of ability parameter (Baker, 1987; Swaminathan & Gifford, 1985). The studies of model fit also indicate that the 3PL model does not necessarily provide a better fit than the 2PL model. For example, Yen (1981) found that the 2PL model fit simulated data as well as the 3PL model when the true model was the 3PL. Haberman (2006) derived a score test to check if the 3PL is better than the 2PL and examined data from a teacher certification test. His results suggested that the gain in data description from use of a 3PL rather than a 2PL was small and the routine use of the 3PL model may not be warranted given the computational difficulty associated with the c parameter. Therefore, in this study, we used the 2PL model for the dichotomous items.

After fitting the models to the simulated data, the correlations and root mean squared differences (RMSD) between the true parameters and estimated parameters were evaluated. Because the true parameters θ_j and γ_{dj} of the general polytomous testlet model were simulated from a $N(0,1)$ standard normal distribution, and during the estimation process, θ_j and γ_{dj} were also fixed to the $N(0,1)$ distribution for identification purpose, the estimated parameters and

true parameters are on the same scale. When estimating the 2PL/GPCM, θ_j was also constrained to have a $N(0,1)$ distribution so that the estimated parameters and true parameters are on the same scale. Similarly, for the real data analyzed in this study (described in the next section), the same $N(0,1)$ constraint was imposed for θ_j and γ_{dj} (general polytomous testlet model) and for θ_j (2PL/GPCM) for all calibration runs. Thus, the parameter estimates from different calibration runs are on the same scale.

Results of Simulation Study

Item Parameter Recovery

Tables 1 and 2 display the correlations and RMSDs between the true and estimated item parameters from the general polytomous testlet model and the 2PL/GPCM. It appears that the item discrimination parameters a_{i1} and item difficulty parameters t_i were well estimated when fitting the general polytomous testlet model to the data. For the polytomous items, because there were only three of them and each of the polytomous items were scored 0, 1, and 2, the correlations and RMSDs for item category parameter d_{iv} (the t_{iv} may be decomposed as $t_{iv} = t_i + d_{iv}$) were not separately calculated. Instead, the parameters t_i for these items were included in the correlation and RMSD calculations for each data set. The average correlation across the two conditions for the discrimination parameters, a_{i1} and a_{i2}, and item difficulty parameters t_i were 0.9680, 0.8608, and 0.9982, respectively. The average RMSD values across the two conditions were reasonably low: 0.0671 for the difficulty parameters, 0.1165 for the a_{i1} parameters, and 0.1148 for the a_{i2} parameters. These correlations and RMSD values appeared to be acceptable compared to those reported in the literature. Similar results were obtained under Conditions 1 and 2 for the a_{i1} and t_i parameters. However, Condition 2 (large testlet effect) produced more accurate estimates for the a_{i2} parameters (higher correlations and lower RMSDs) than Condition 1.

When the 2PL/GPCM was fit to the simulated data (local dependence was ignored), under Condition 1, the correlations and RMSD results for the a_{i1} and t_i parameters were similar to or better than those obtained when fitting the general polytomous testlet model. Under condition 2 (larger

testlet effect), although similar results were obtained for the slope parameters, the RMSDs for the difficulty parameters were higher than those produced by the general polytomous testlet model.

Table 1

Correlation Between True Item Parameters and Estimated Parameters from the General Polytomous Testlet Model and the 2PL/GPCM

		General testlet model			2PL/GPCM	
Condition	Data set	a_{i1}	a_{i2}	t_i	a_{i1}	t_i
Small testlet effect	Data 1	0.9837	0.8412	0.9980	0.9818	0.9977
	Data 2	0.9802	0.8423	0.9978	0.9796	0.9970
	Data 3	0.9586	0.8612	0.9984	0.9516	0.9978
	Data 4	0.9692	0.8274	0.9993	0.9686	0.9990
	Data 5	0.9499	0.7143	0.9989	0.9496	0.9990
	Data 6	0.9499	0.8400	0.9985	0.9365	0.9974
	Data 7	0.9736	0.8183	0.9982	0.9766	0.9982
	Data 8	0.9793	0.9204	0.9983	0.9784	0.9980
	Data 9	0.9704	0.6815	0.9966	0.9763	0.9964
	Data 10	0.9587	0.8184	0.9983	0.9594	0.9982
	Mean	0.9674	0.8165	0.9982	0.9658	0.9979
Larger testlet effect	Data 1	0.9753	0.8718	0.9979	0.9656	0.9980
	Data 2	0.9550	0.8927	0.9967	0.9489	0.9973
	Data 3	0.9835	0.9243	0.9985	0.9730	0.9973
	Data 4	0.9797	0.9421	0.9979	0.9780	0.9974
	Data 5	0.9813	0.9368	0.9983	0.9586	0.9977
	Data 6	0.9684	0.9023	0.9978	0.9568	0.9967
	Data 7	0.9777	0.8925	0.9989	0.9764	0.9980
	Data 8	0.9450	0.9036	0.9984	0.9489	0.9967
	Data 9	0.9538	0.8712	0.9985	0.9586	0.9982
	Data 10	0.9666	0.9133	0.9981	0.9496	0.9982
	Mean	0.9686	0.9051	0.9981	0.9614	0.9976
All	Overall mean	0.9680	0.8608	0.9982	0.9636	0.9977

Table 2

RMSD Between True Item Parameters and Estimated Parameters from the General Polytomous Testlet Model and the 2PL/GPCM

		General testlet model			2PL/GPCM	
Condition	Data set	a_{i1}	a_{i2}	t_i	a_{i1}	t_i
Small testlet effect	Data 1	0.1430	0.1339	0.0554	0.0690	0.0568
	Data 2	0.1306	0.1482	0.0709	0.0684	0.0914
	Data 3	0.0878	0.1110	0.0635	0.0782	0.0813
	Data 4	0.1296	0.1215	0.0440	0.0682	0.0668
	Data 5	0.1436	0.1554	0.0631	0.0722	0.0653
	Data 6	0.1350	0.1448	0.0638	0.0852	0.0873
	Data 7	0.1419	0.1288	0.0710	0.0780	0.0679
	Data 8	0.1073	0.1287	0.0818	0.0692	0.0671
	Data 9	0.1240	0.1724	0.0741	0.0787	0.0661
	Data 10	0.1260	0.1287	0.0596	0.0658	0.0582
	Mean	0.1269	0.1373	0.0647	0.0733	0.0708
Larger testlet effect	Data 1	0.1232	0.0971	0.0525	0.1070	0.1018
	Data 2	0.0857	0.0913	0.0705	0.0949	0.0928
	Data 3	0.1207	0.0892	0.0672	0.1063	0.1307
	Data 4	0.0936	0.0876	0.0788	0.1021	0.0864
	Data 5	0.0731	0.0786	0.0653	0.1137	0.1042
	Data 6	0.1101	0.1089	0.0632	0.0983	0.1227
	Data 7	0.1467	0.0908	0.0680	0.1079	0.1456
	Data 8	0.1027	0.1012	0.0692	0.1025	0.1087
	Data 9	0.1173	0.0920	0.0819	0.0921	0.1216
	Data 10	0.0876	0.0865	0.0781	0.1103	0.1029
	Mean	0.1061	0.0923	0.0695	0.1035	0.1117
All	Overall mean	0.1165	0.1148	0.0671	0.0884	0.0913

After computing the mean differences between the true and estimated item parameters, a slight underestimation was found for the a_{i1} parameters when the general polytomous testlet model was fit. The average mean difference across the 20 data sets was -0.10.

Model Fit

The likelihood ratio test, Akaike's information criterion (AIC; Akaike, 1974) and Bayesian information criterion (BIC; Schwarz, 1978) were calculated to compare the fit of the two models. As expected, for all simulated data sets, both the likelihood ratio test and AIC gave consistent results (not shown in this paper) indicating that the general polytomous testlet model fit better than the 2PL/GPCM. This is because the data were generated using the general polytomous testlet model. For 7 of the 20 data sets, BIC preferred the 2PL/GPCM. This is probably because BIC gives a higher penalty if the number of parameters is large and thus tends to choose models with fewer parameters than the AIC (Lin and Dayton, 1997). These results suggested that the AIC and likelihood ratio test would be more effective than the BIC in identifying the true model for these data.

Test Information

Previous studies (e. g., Wang et al., 2002) suggested that the presence of testlet effects could affect test information. Test information provides crucial information about the precision of examinee ability estimation at different trait levels. For polytomously scored item response models, the item information function proposed by Samejima (1974) is

$$I_i(\theta) = a_i^2 \sum_{c=1}^{m_i} \left[T_c - \overline{T}_i(\theta)\right]^2 P_{ic}(\theta), \qquad (3)$$

where $\overline{T}_i(\theta) = \sum_{c=1}^{m_i} T_c P_{ic}(\theta)$, and T_c is a scoring function for item score category $c = 1, 2, 3, \cdots, m_j$.

The test information is simply the sum of item information across all items on the test. Using Equation 3, we can draw test information curves estimated from both the general polytomous testlet model and the 2PL/GPCM for each data set. The test information was computed for 100 equally spaced points of θ between -4 and 4. For these data, the scoring function T_c takes on values of 0, 1, and 2. To calculate the test information curve based on general polytomous testlet model, the item parameters estimated from the simulated data were used to calculate the probability $P_{ic}(\theta)$ and the test information.

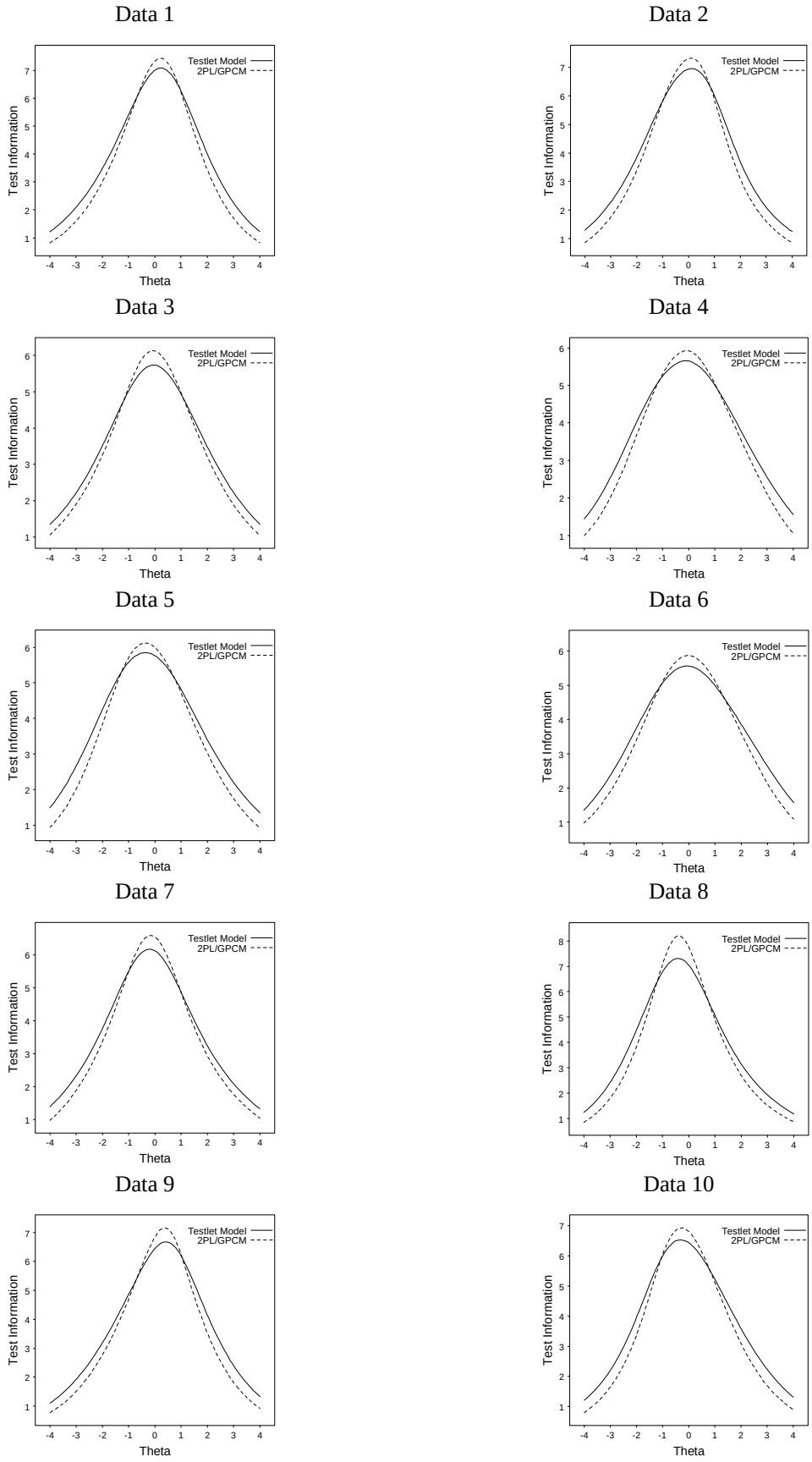

Figure 1. **Test information for simulated data sets 1–10 under Condition 1**

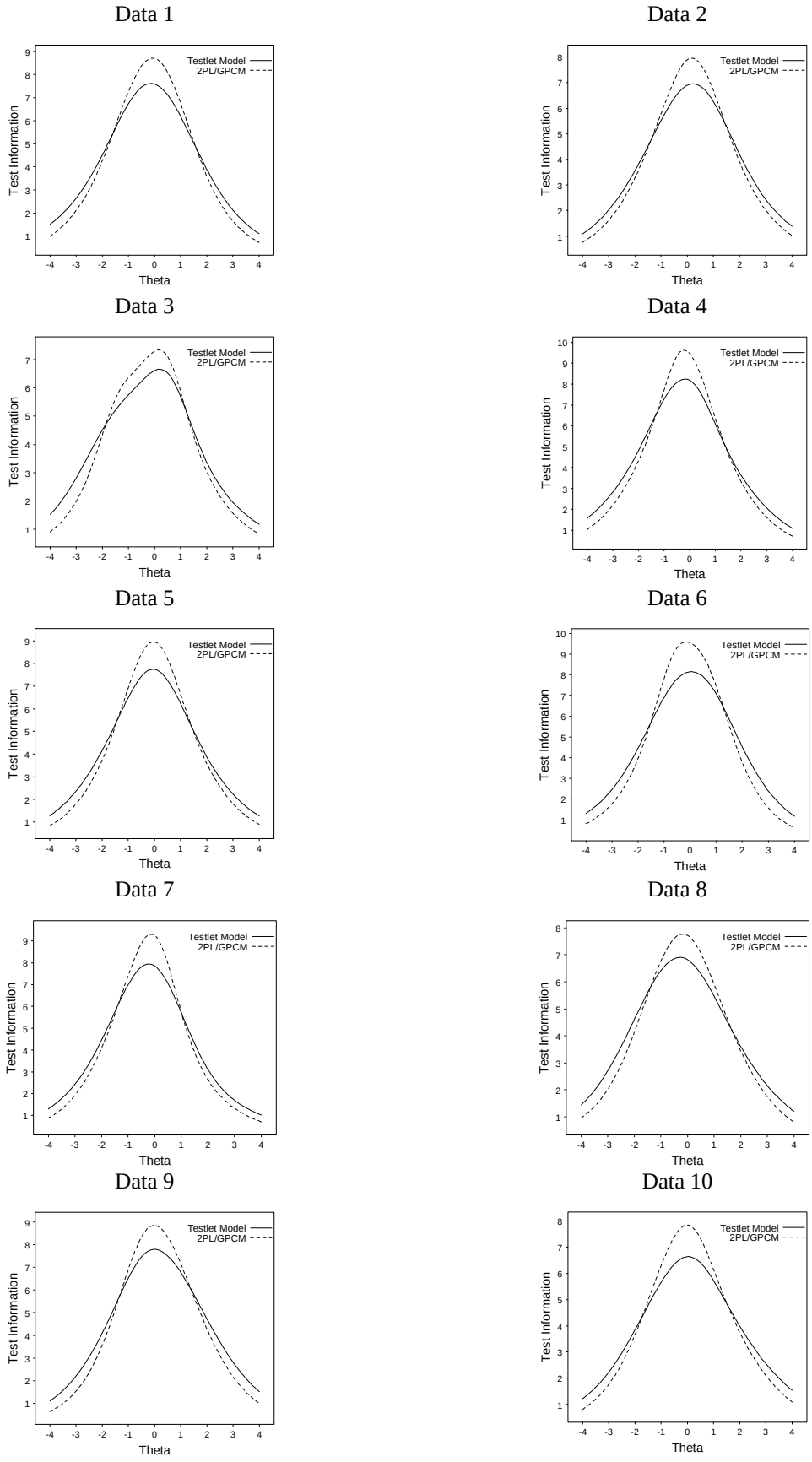

Figure 2 **. Test information for simulated data sets 1–10 under Condition 2**

At each θ point, 1000 γ_d were drawn from a standard normal distribution. The test information plotted is the average information over γ_d s. These test information curves are shown in Figures 1 and 2. Under both Condition 1 and Condition 2, compared to the test information derived from general polytomous testlet model in which the local dependence among testlet items is taken into account, the test information estimated from the 2PL/GPCM was higher at the middle range of ability levels and lower at extreme ability levels. The overestimation of test information was larger under the larger testlet effect condition (Condition 2).

Table 3

Comparison Between Passage-based Reliability and Item-based Reliability for the Simulated Data

Condition	Data set	Item-based reliability	Passage-based reliability	Proportion of decrease
Small testlet effect	Data 1	0.8692	0.8206	5.6%
	Data 2	0.8713	0.8156	6.4%
	Data 3	0.8411	0.7830	6.9%
	Data 4	0.8559	0.7933	7.3%
	Data 5	0.8616	0.8204	4.8%
	Data 6	0.8554	0.8008	6.4%
	Data 7	0.8506	0.7956	6.5%
	Data 8	0.8649	0.8085	6.5%
	Data 9	0.8623	0.7989	7.4%
	Data 10	0.8676	0.8201	5.5%
	Mean	0.8600	0.8057	6.3%
Larger testlet effect	Data 1	0.8789	0.6916	21.3%
	Data 2	0.8670	0.7003	19.2%
	Data 3	0.8638	0.7094	17.9%
	Data 4	0.8810	0.7179	18.5%
	Data 5	0.8780	0.6973	20.6%
	Data 6	0.8884	0.7219	18.7%
	Data 7	0.8697	0.7096	18.4%

Condition	Data set	Item-based reliability	Passage-based reliability	Proportion of decrease
Larger testlet effect (*continued*)	Data 8	0.8715	0.7114	18.4%
	Data 9	0.8834	0.7331	17.0%
	Data 10	0.8641	0.6758	21.8%
	Mean	0.8746	0.7069	19.2%
All	Overall mean	0.8673	0.7563	12.7%

We also computed the passage-based Cronbach's alpha for each of the simulated data sets, that is, summing the scored responses across items within a testlet and then using the testlet score in calculating Cronbach's alpha in order to eliminate the effect of local item dependence within a testlet (Sireci et al., 1991; Wainer & Thissen, 1996). The passage-based as well as item-based reliability estimates are provided in Table 3. Under the small testlet effect condition, the proportion of decrease in reliability estimates by computing the passage-based alpha ranged from 4.8% to 7.4%. Large decreases in reliability estimates were observed for Condition 2, and the proportion of decrease ranged from 17.0% to 21.8%. These results are consistent with what we found in the test information analyses, suggesting that the reliability was overestimated when local dependence was present.

Application to a Large-scale English Language Test

Data

The real data for this study came from the reading sections of six operational test forms of a large-scale English language test administered between 2006 and 2007. The focus of this part of the study was on the application of the proposed model to operational data sets. For each of the six forms, a random sample of 2,000 examinees was selected. Each of the six reading tests contained 3 passages, with 13–14 items per passage. The last item in each passage was polytomously scored (0, 1, 2, 3), while the remaining items were dichotomously scored (0, 1).

The general polytomous testlet model was first fit to each of the six reading data sets. Next, a standard 2PL/GPCM model was fit, i.e., local dependence was not taken into account. The likelihood ratio test, AIC and BIC were calculated to compare the two models and determine which model provided a better fit to the data. In addition, item discrimination and item difficulty parameter estimates obtained from the two models were compared to see if any differences existed.

Correlations and RMSDs between the two sets of parameter estimates were also computed. Finally, test information curves based on the two models were plotted and compared.

Results

Model Fit

Table 4 shows the results of the likelihood ratio test, the AIC and BIC indices for the six real data sets. These data sets will be referred to as Data A, Data B, …, Data F, and their associated tests will be referred to as Test A, Test B, …, Test F. For Data A and F, the likelihood ratio test, AIC, and BIC all suggested that the general polytomous testlet model did not fit better than the simple 2PL/GPCM. For the remaining four datasets: Data B, Data C, Data D, and Data E, the G^2 was significant across all four data sets. The AIC suggested that general polytomous testlet model fit data sets C and D slightly better than the 2PL/GPCM, but the simpler 2PL/GPCM model fit data sets B and E better than the general polytomous testlet model. The BIC preferred simpler model over the complex model for all the six data sets.

Table 4

Goodness-of-fit of the General Polytomous Testlet Model versus the 2PL (GPCM)

Data	Model	-2logL	G^2	df	AIC	BIC
	Testlet model	96002			96256	96967
Data A	GPCM	96038	36	41	96210	96692
	Testlet model	99575			99833	100556
Data B	GPCM	99647	72*	42	99821	100309
	Testlet model	99434			99692	100415
Data C	GPCM	99531	97*	42	99705	100192
	Testlet model	93036			93294	94016
Data D	GPCM	93122	86*	42	93296	93783
	Testlet model	99784			100042	100764
Data E	GPCM	99853	69*	42	100027	100514
	Testlet model	88135			88393	89115
Data F	GPCM	88184	49	42	88358	88845

* $p < .05$

To check which items were affected by the secondary dimensions (local dependence), we examined the estimated a_{i2} parameters. As an example, Table 5 provides estimated item discrimination parameters and their standard errors for one data set. Several items had large a_{i2} parameters; for example, item 12 in passage 1 (0.89), items 22 and 24 in passage 2 (0.88, 1.00), and item 40 in passage 3 (1.19). Apparently, these items were affected by the secondary dimensions introduced by the testlets. Overall, it appeared that the general polytomous testlet model did not describe these data much better than the 2PL/GPCM. As will be described shortly, a simpler IRT model was also fit to these data to check possible presence of testlet effect.

Table 5

Estimated Item Discrimination Parameters for Data C

Passage	Item	$\hat{a}_{i1}$	SE	$\hat{a}_{i2}$	SE
1	1	0.85	0.06	0.15	0.09
	2	1.18	0.07	0.00	-
	3	0.73	0.06	0.42	0.10
	4	0.39	0.05	0.08	0.09
	5	0.90	0.07	0.12	0.11
	6	1.60	0.12	0.76	0.15
	7	1.04	0.09	0.70	0.13
	8	0.90	0.06	0.04	0.10
	9	0.66	0.06	0.59	0.11
	10	1.16	0.08	0.39	0.12
	11	0.83	0.06	0.31	0.10
	12	0.86	0.07	0.89	0.14
	13	0.68	0.05	0.46	0.10
	14	0.73	0.05	0.59	0.09

Table continues

Passage	Item	$\hat{a}_{i1}$	SE	$\hat{a}_{i2}$	SE
2	15	0.46	0.05	0.07	0.09
	16	0.88	0.06	0.52	0.11
	17	0.63	0.05	0.36	0.10
	18	0.91	0.06	0.50	0.11
	19	0.95	0.07	0.46	0.11
	20	0.87	0.06	0.61	0.11
	21	1.13	0.07	0.58	0.11
	22	1.74	0.13	0.88	0.16
	23	0.72	0.06	0.52	0.11
	24	2.22	0.22	1.00	0.21
	25	1.01	0.07	0.75	0.12
	26	0.78	0.06	0.44	0.11
	27	0.67	0.05	0.12	0.10
	28	1.06	0.06	0.32	0.08
3	29	0.60	0.05	0.00	-
	30	1.11	0.08	0.22	0.09
	31	1.21	0.08	0.22	0.09
	32	0.99	0.07	0.10	0.09
	33	1.25	0.08	0.34	0.09
	34	0.45	0.06	0.08	0.10
	35	1.04	0.08	0.49	0.10
	36	1.04	0.07	0.39	0.09
	37	1.11	0.08	0.71	0.11
	38	1.00	0.06	0.59	0.09
	39	0.92	0.06	0.45	0.09
	40	1.09	0.08	1.19	0.15
	41	1.17	0.07	0.75	0.11
	42	0.77	0.05	0.77	0.10

Item Parameter Estimates

The correlations and RMSDs between the estimated item parameters from the general polytomous testlet model and the 2PL/GPCM for the real data sets are shown in Tables 6 and 7. As can be seen, the item discrimination parameter (with respect to θ) and difficulty parameters estimated from the two models were highly correlated, with an average correlation of 0.9893 for the discrimination parameters and an average correlation of 0.9978 for the difficulty parameters. Overall, the RMSDs between the item parameters estimated from the two models were small. The mean differences between the two sets of item parameters were also calculated to see if one model produced higher or lower parameters than the other model. While the mean differences for the difficulty parameters were very small, slightly higher item discrimination parameters were found for the 2PL/GPCM. For these data sets, the item parameter estimates appeared to be little affected by using the 2PL/GPCM model that ignored local item dependence.

Table 6

Correlation Between Estimated Item Parameters from the General Polytomous Testlet Model and the 2PL/GPCM for the Real Data Sets

Data	Slope (a)	Difficulty(t)
Data A	0.9877	0.9987
Data B	0.9871	0.9970
Data C	0.9958	0.9993
Data D	0.9929	0.9976
Data E	0.9897	0.9987
Data F	0.9823	0.9954
Mean	0.9893	0.9978

Test Information

The test information curves for the real data sets are shown in Figure 3. For all six data sets, higher test information at middle range of ability levels were found using the 2PL/GPCM, in which local dependence was ignored. Table 8 shows the passage-based as well as item-based reliability estimates for real data sets. The proportion of decrease in reliability estimates by computing the passage-based alpha ranged from 2.3% to 4.9%, suggesting that the reliability was overestimated for these data if local dependence is ignored. These results are consistent with what we found in the

test information analyses. For Data A and Data F, contrary to the results based on the likelihood ratio test, AIC, and BIC, the reliability results showed that there were testlet effects in these data. This inconsistency is mainly due to the fact that the general polytomous testlet model is a more complex model, with many parameters to be estimated. The model fit results suggest that the improvement in model fit is not worth the cost of estimating the additional parameters, but it does not necessarily mean the absence of testlet effects in these data. In the next section, we examine the fit of a multidimensional IRT model with simple structure to these data.

Table 7

RMSD and Mean Difference Between Estimated Item Parameters from the General Polytomous Testlet Model and the 2PL/GPCM for the Real Data Sets

	Slope (a)		Difficulty (t)	
Data	RMSD	Mean difference	RMSD	Mean difference
Data A	0.1140	-0.0866	0.0713	0.0485
Data B	0.1416	-0.1287	0.0936	0.0640
Data C	0.1339	-0.1210	0.1132	0.1070
Data D	0.1357	-0.1210	0.0748	-0.0278
Data E	0.1291	-0.1150	0.0705	0.0544
Data F	0.1965	-0.1658	0.1171	0.0204
Mean	0.1418	-0.1230	0.0901	0.0444

Table 8

Comparison Between Passage-based Reliability and Item-based Reliability for the Real Data Sets

Data set	Item-based reliability	Passage-based reliability	Proportion of decrease
Data A	0.8599	0.8179	4.9%
Data B	0.8733	0.8415	3.6%
Data C	0.8877	0.8605	3.1%
Data D	0.8827	0.8624	2.3%
Data E	0.8716	0.8476	2.8%
Data F	0.8786	0.8451	3.8%

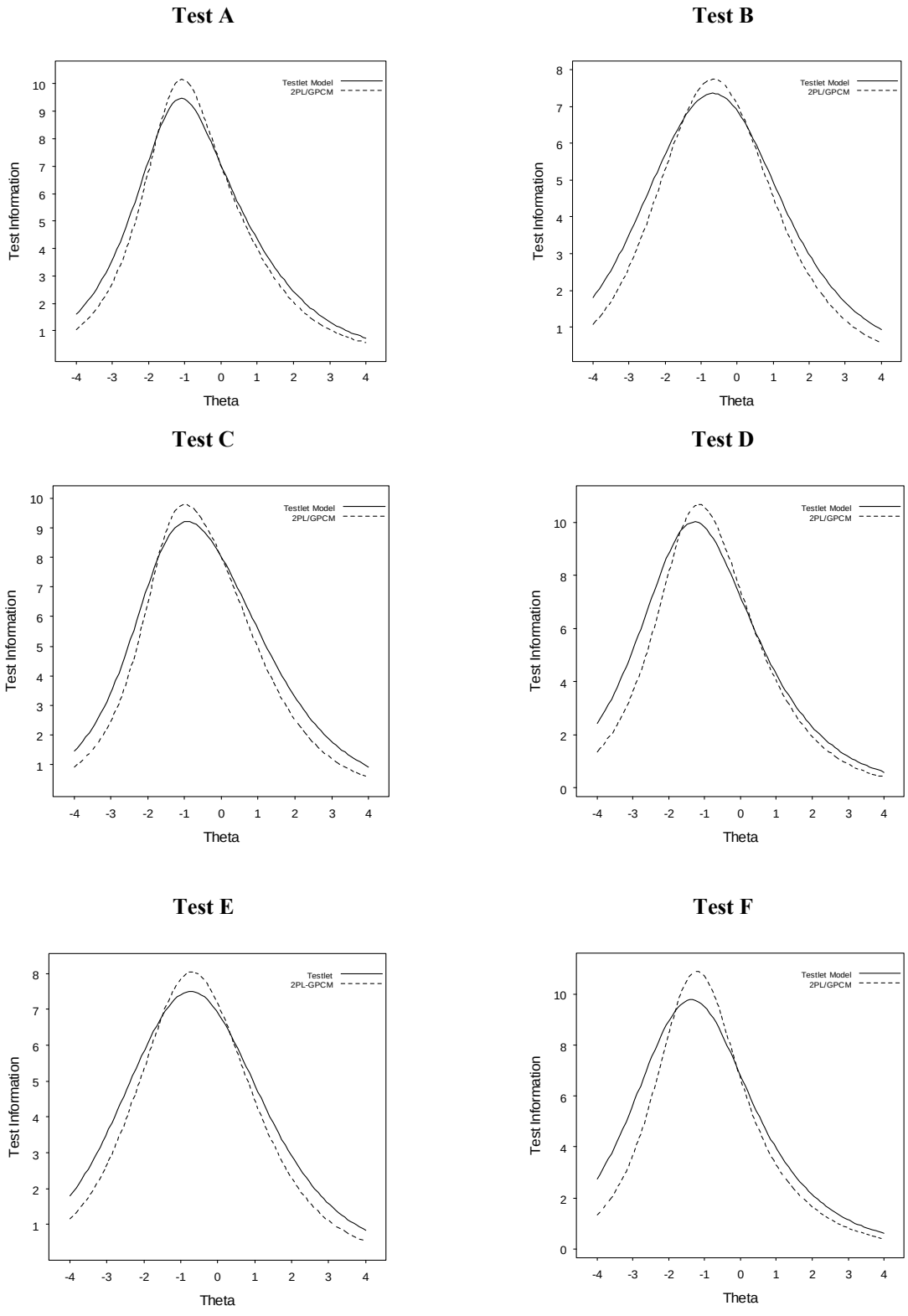

Figure 3. **Test information for the reading section of Tests A–F**

A Multidimensional IRT Model with Simple Structure

Although the general polytomous testlet model accounts for local dependence among testlet items, it has many parameters to be estimated. For example, the English language test used in this study had 3 sets (reading passages) and each set contained 13-14 items, and thus the number of parameters to be estimated for each test was 127-129. For some data sets, the fit of general polytomous testlet model may not be good due to its complexity. The following is a simpler multidimensional IRT model with simple structure (MIRT-SS), meaning that each item has only one non-zero loading on the latent traits:

$$P_{ijk} = \frac{\exp\left[\Sigma_{v=0}^{k}\left(a_i\theta_{d(i)j} - t_{iv}\right)\right]}{\Sigma_{c=0}^{m_i}\exp\left[\Sigma_{v=0}^{c}\left(a_i\theta_{d(i)j} - t_{iv}\right)\right]}, \tag{4}$$

where P_{ijk} is the probability of scoring in category k of the $m_i + 1$ score categories of item i; $\theta_{d(i)j}$ is the only latent trait related to item i and follows a multivariate standard normal distribution with a correlation matrix $\boldsymbol{\Sigma}$, that is, $\theta_{d(i)j} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$; t_{iv} and a_i have the same interpretations as those in Equation 2; and $\Sigma_{v=0}^{0}\left(a_i\theta_{d(i)j} - t_{iv}\right) \equiv 0$.

This model assumes that there is one latent trait for each testlet (a total of 3 latent traits for these data), and the latent traits are correlated. As can be seen in Equation 4, the number of parameters in this model is significantly reduced when compared to the general polytomous testlet model. If a testlet effect is present, it is expected that the multidimensional IRT model with simple structure (MIRT-SS) would fit better than the 2PL/GPCM, as it takes into account the multidimensionality introduced by the testlets. Compared with the general polytomous testlet model, the MIRT-SS model has far fewer item parameters, and it captures the testlet effect through a separate ability dimension for each testlet, while the ability dimensions are allowed to be correlated. The MIRT-SS model was estimated using the SAS NLMIXED procedure. Appendix A gives a sample SAS code for fitting this model.

The MIRT-SS model was fit to each of the six reading data sets. The estimated latent traits were highly correlated, ranging from 0.87 to 0.93. The likelihood ratio test, AIC and BIC were calculated to evaluate the fit of the models. Table 9 gives the results of the likelihood ratio test for comparing the MIRT-SS model with the 2PL/GPCM. For all six datasets, G^2 was significant, indicating that the MIRT-SS fit the data better than the 2PL/GPCM model.

Table 10 shows the AIC and BIC indices for the three models: 2PL/GPCM, the general polytomous testlet model, and the MIRT-SS model. For comparison purposes, the AIC and BIC for the general polytomous testlet model shown in Table 4 were repeated in Table 10. As can be seen, both AIC and BIC indicated that the MIRT-SS model fit the data better than the 2PL/GPCM and the general testlet model for all data sets except Data D, for which the fit of the three models was similar. These results suggested the presence of local dependence in these data sets. Note that the comparison of the fit between the MIRT-SS model and the 2PL/GPCM suggested that for Data A and F, testlet effects were present, which was consistent with the previous reliability analysis results.

Table 9

Likelihood Ratio Test for Comparing the MIRT-SS Model with the 2PL (GPCM) Model

Data	Model	-2logL	G^2	df
Data A	MIRT-SS	95930	108*	3
	GPCM	96038		
Data B	MIRT-SS	99569	78*	3
	GPCM	99647		
Data C	MIRT-SS	99423	108*	3
	GPCM	99531		
Data D	MIRT-SS	93110	12*	3
	GPCM	93122		
Data E	MIRT-SS	99813	40*	3
	GPCM	99853		
Data F	MIRT-SS	88140	44*	3
	GPCM	88184		

* $p < .05$

Table 10

AIC and BIC Indices for the MIRT-SS Model, the Testlet Model, and the 2PL/GPCM Model

Data	Model	AIC	BIC
Data A	MIRT-SS	96108	96607
	Testlet model	96256	96967
	GPCM	96210	96692
Data B	MIRT-SS	99749	100253
	Testlet model	99833	100556
	GPCM	99821	100309
Data C	MIRT-SS	99603	100107
	Testlet model	99692	100415
	GPCM	99705	100192
Data D	MIRT-SS	93290	93794
	Testlet model	93294	94016
	GPCM	93296	93783
Data E	MIRT-SS	99993	100497
	Testlet model	100042	100764
	GPCM	100027	100514
Data F	MIRT-SS	88320	88824
	Testlet model	88393	89115
	GPCM	88358	88845

Discussion

In this study, the application of a general polytomous testlet model to a large-scale English language assessment was investigated. The general polytomous testlet model can be applied to tests composed of both dichotomous and polytomous testlet items. The model extends a previous bifactor analysis approach to polytomous testlet items by using a generalized partial credit model. This model not only takes into account local dependence within the testlets but also provides more information about how items within a testlet are influenced by the testlet factor. However, one drawback of this model is that more item parameters need to be estimated, and thus the fit of the model may not be good. A simpler multidimensional IRT model with simple structure was also applied to the real data in this study, and the results suggested a better fit than that of the general polytomous testlet model.

The analyses of the six real data sets indicated the presence of local dependence in these data, which seemed to have a small impact on item parameter estimates and a relatively larger impact on test information and reliability estimates. In these analyses, we applied the likelihood ratio test, AIC and BIC to compare the fit of three alternative models to these data. Other fit statistics may also be useful in investigating the local item dependence present in these data. For example, Orlando and Thissen's (2000) Q_1 and G^2 tests can be extended to evaluate item fit under the current model. Analyses of item pair odds ratios as well as residuals would also be helpful in determining model-data fit. These additional approaches should be considered in future studies. A simulation study was conducted to evaluate the parameter recovery of the general polytomous testlet model. The results suggested that the recovery of the true item parameters was good, as indicated by the high correlations and low RMSDs between the true parameters and estimated parameters. The impact of local dependence on item parameter estimation was also investigated by fitting a 2PL/GPCM to the data. It appeared that under the small testlet effect condition (Condition 1) the estimated item parameters were accurate. However, if a larger testlet effect was present (Condition 2), the RMSD between the true and estimated item parameters became larger. This study also demonstrated the flexibility of SAS NLMIXED in fitting IRT models. Although the model was easily implemented using the NLMIXED procedure, the computational time was quite long (about 20–35 hours), which limits its use in extensive data analyses. Thus, alternative software for implementing this model needs to be explored in the future. Given that it is time-consuming to fit the general polytomous testlet model, some relatively simple methods to detect local dependence and measure the magnitude

of the testlet effect may be used before applying the model. For example, for dichotomous items, the usual Q_3 (Yen, 1984) statistic can be used to assess local dependence. Alternatively, one can compare item-based reliability estimate with testlet-based reliability estimate (Sireci et al., 1991; Wainer & Thissen, 1996), which was used in this study. This method can be applied to both dichotomous and polytomous items.

Finally, more studies on the general polytomous testlet model are needed. In this study we analyzed the reading sections of six operational large scale language tests. It would be interesting to examine the impact of local dependence on the pretest (equating) items, because the quality of item parameter estimates plays an important role in IRT based equating. In addition, the utility of the proposed model can be studied using data from other testing programs that also contain testlet items.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723.

Baker, F. B. (1987). Methodology review: Item parameter estimation under the one-, two-, and three-parameter logistic models. *Applied Psychological Measurement,* 12, 111–141.

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

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer-Verlag.

Gibbons, R. D., & Hedeker, D. R. (1992). Full information bi-factor analysis. *Psychometrika*, 57, 423–436.

Haberman S. J. (2006). *An elementary test of the normal 2PL model against the normal 3PL alternative.* (ETS Research Rep. RR-06-14). Princeton, NJ: ETS.

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

Lin, T. H., & Dayton, C. M. (1997). Model selection information criteria for non-nested latent class models. *Journal of Educational and Behavioral Statistics, 22*(3), 249–264.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

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

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*, 50–65.

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods*, 8, 185–205.

Samejima, F. (1974). Normal ogive model on the continuous response level in the multidimensional latent space. *Psychometrika, 39,* 111–121.

SAS Institute. (1999). The NLMIXED procedure [Computer software]. Cary, NC: Author.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461–464.

Sheu, C.-F., Chen, C.-T., Su, Y.-H., & Wang, W.-C. (2005). Using SAS PROC NLMIXED to fit item response theory models. *Behavior Research Methods,* 37, 202–218.

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

Swaminathan, H., & Gifford, J. A. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika, 50,* 349–364.

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 adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 245–270). Boston, MA: Kluwer-Nijhoff.

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

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

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5,* 245–262.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8,* 125–145.

Appendix

An Example of SAS Code Used for Estimating the General Polytomous Testlet Model, the 2PL/GPCM Model, and the Multidimensional IRT Model with Simple Structure

```
*Fit the general polytomous testlet model to the data using NLMIXED;
title 'The General Polytomous Testlet Model';
proc nlmixed data=DataB_sample_t method=gauss noad technique=quanew qpoints=4;
bounds a1_1-a1_42 >0, a2_1-a2_42 >0;
parms a1_1-a1_42=1 a2_1-a2_42=0.5 b1-b42=0 d14_1=0 d28_1=0 d42_1=0;
beta = b1   *i1   +
b2    *i2   +
b3    *i3   +
b4    *i4   +
b5    *i5   +
b6    *i6   +
b7    *i7   +
b8    *i8   +
b9    *i9   +
b10   *i10  +
b11   *i11  +
b12   *i12  +
b13   *i13  +
b14   *i14  +
b15   *i15  +
b16   *i16  +
b17   *i17  +
b18   *i18  +
b19   *i19  +
b20   *i20  +
b21   *i21  +
b22   *i22  +
b23   *i23  +
b24   *i24  +
b25   *i25  +
b26   *i26  +
b27   *i27  +
b28   *i28  +
b29   *i29  +
b30   *i30  +
b31   *i31  +
b32   *i32  +
b33   *i33  +
b34   *i34  +
b35   *i35  +
b36   *i36  +
b37   *i37  +
b38   *i38  +
b39   *i39  +
b40   *i40  +
b41   *i41    +
b42   *i42
;
slope1=
a1_1  *i1   +
```

```
a1_2  *i2   +
a1_3  *i3   +
a1_4  *i4   +
a1_5  *i5   +
a1_6  *i6   +
a1_7  *i7   +
a1_8  *i8   +
a1_9  *i9   +
a1_10 *i10  +
a1_11 *i11  +
a1_12 *i12  +
a1_13 *i13  +
a1_14 *i14  +
a1_15 *i15  +
a1_16 *i16  +
a1_17 *i17  +
a1_18 *i18  +
a1_19 *i19  +
a1_20 *i20  +
a1_21 *i21  +
a1_22 *i22  +
a1_23 *i23  +
a1_24 *i24  +
a1_25 *i25  +
a1_26 *i26  +
a1_27 *i27  +
a1_28 *i28  +
a1_29 *i29  +
a1_30 *i30  +
a1_31 *i31  +
a1_32 *i32  +
a1_33 *i33  +
a1_34 *i34  +
a1_35 *i35  +
a1_36 *i36  +
a1_37 *i37  +
a1_38 *i38  +
a1_39 *i39  +
a1_40 *i40  +
a1_41 *i41   +
a1_42 *i42
;

slope2=
a2_1  *i1   +
a2_2  *i2   +
a2_3  *i3   +
a2_4  *i4   +
a2_5  *i5   +
a2_6  *i6   +
a2_7  *i7   +
a2_8  *i8   +
a2_9  *i9   +
a2_10 *i10  +
a2_11 *i11  +
a2_12 *i12  +
a2_13 *i13  +
```

```
a2_14 *i14  +
a2_15 *i15  +
a2_16 *i16  +
a2_17 *i17  +
a2_18 *i18  +
a2_19 *i19  +
a2_20 *i20  +
a2_21 *i21  +
a2_22 *i22  +
a2_23 *i23  +
a2_24 *i24  +
a2_25 *i25  +
a2_26 *i26  +
a2_27 *i27  +
a2_28 *i28  +
a2_29 *i29  +
a2_30 *i30  +
a2_31 *i31  +
a2_32 *i32  +
a2_33 *i33  +
a2_34 *i34  +
a2_35 *i35  +
a2_36 *i36  +
a2_37 *i37  +
a2_38 *i38  +
a2_39 *i39  +
a2_40 *i40  +
a2_41 *i41    +
a2_42 *i42
;
d1= d14_1*i14 + d28_1*i28 + d42_1*i42;
eta1= (slope1*theta - beta +
slope2*gamma1*p1+slope2*gamma2*p2+slope2*gamma3*p3)*(i1     +
i2    +
i3    +
i4    +
i5    +
i6    +
i7    +
i8    +
i9    +
i10   +
i11   +
i12   +
i13   +
i15   +
i16   +
i17   +
i18   +
i19   +
i20   +
i21   +
i22   +
i23   +
i24   +
i25   +
i26   +
```

```
i27   +
i29   +
i30   +
i31   +
i32   +
i33   +
i34   +
i35   +
i36   +
i37   +
i38   +
i39   +
i40   +
i41)
+(slope1*theta - beta-d1 +
slope2*gamma1*p1+slope2*gamma2*p2+slope2*gamma3*p3)*(i14+
i28+i42);
num1 = exp(eta1);
num2 = exp(2*eta1+2*d1)*(i14+i28+i42);
denom=1+num1+num2;
if (score1=0) then prob=1/denom;
else if (score1=1) then prob=num1/denom;
else if (score1=2) then prob=num2/denom;
if(prob>1E-8) then ll=log(prob);
else ll=-1E100;
model score1 ~ general(ll);
random theta gamma1 gamma2 gamma3 ~ Normal ([0,0,0,0],[1,0,1,0,0,1,0,0,0,1])
subject= person;
ods output ParameterEstimates=DataB_itempar_m1;
run;



*Fit the 2PL (or GPCM) model to the same data using NLMIXED;

title 'The 2PL(GPCM) Model';
proc nlmixed data=DataB_sample_t method=gauss noad technique=quanew qpoints=20;
bounds a1_1-a1_42 >0;
parms a1_1-a1_42=1 b1-b42=0 d14_1=0 d28_1=0 d42_1=0;
beta = b1   *i1   +
b2    *i2   +
b3    *i3   +
b4    *i4   +
b5    *i5   +
b6    *i6   +
b7    *i7   +
b8    *i8   +
b9    *i9   +
b10   *i10  +
b11   *i11  +
b12   *i12  +
b13   *i13  +
b14   *i14  +
b15   *i15  +
b16   *i16  +
b17   *i17  +
b18   *i18  +
```

```
b19   *i19  +
b20   *i20  +
b21   *i21  +
b22   *i22  +
b23   *i23  +
b24   *i24  +
b25   *i25  +
b26   *i26  +
b27   *i27  +
b28   *i28  +
b29   *i29  +
b30   *i30  +
b31   *i31  +
b32   *i32  +
b33   *i33  +
b34   *i34  +
b35   *i35  +
b36   *i36  +
b37   *i37  +
b38   *i38  +
b39   *i39  +
b40   *i40  +
b41   *i41    +
b42   *i42;
slope1=
a1_1  *i1   +
a1_2  *i2   +
a1_3  *i3   +
a1_4  *i4   +
a1_5  *i5   +
a1_6  *i6   +
a1_7  *i7   +
a1_8  *i8   +
a1_9  *i9   +
a1_10 *i10  +
a1_11 *i11  +
a1_12 *i12  +
a1_13 *i13  +
a1_14 *i14  +
a1_15 *i15  +
a1_16 *i16  +
a1_17 *i17  +
a1_18 *i18  +
a1_19 *i19  +
a1_20 *i20  +
a1_21 *i21  +
a1_22 *i22  +
a1_23 *i23  +
a1_24 *i24  +
a1_25 *i25  +
a1_26 *i26  +
a1_27 *i27  +
a1_28 *i28  +
a1_29 *i29  +
a1_30 *i30  +
a1_31 *i31  +
a1_32 *i32  +
```

```
a1_33 *i33  +
a1_34 *i34  +
a1_35 *i35  +
a1_36 *i36  +
a1_37 *i37  +
a1_38 *i38  +
a1_39 *i39  +
a1_40 *i40  +
a1_41 *i41  +
a1_42 *i42;


d1= d14_1*i14 + d28_1*i28 + d42_1*i42;

eta1= (slope1*theta - beta )*(i1    +
i2    +
i3    +
i4    +
i5    +
i6    +
i7    +
i8    +
i9    +
i10   +
i11   +
i12   +
i13   +
i15   +
i16   +
i17   +
i18   +
i19   +
i20   +
i21   +
i22   +
i23   +
i24   +
i25   +
i26   +
i27   +
i29   +
i30   +
i31   +
i32   +
i33   +
i34   +
i35   +
i36   +
i37   +
i38   +
i39   +
i40   +
i41
)
+(slope1*theta - beta-d1)*(i14+
i28+i42);
num1 = exp(eta1);
```

```
num2 = exp(2*eta1+2*d1)*(i14+i28+i42);
denom=1+num1+num2;
if (score1=0) then prob=1/denom;
else if (score1=1) then prob=num1/denom;
else if (score1=2) then prob=num2/denom;
if(prob>1E-8) then ll=log(prob);
else ll=-1E100;
model score1 ~ general(ll);
random theta ~ Normal (0,1)
subject= person;
ods output ParameterEstimates=DataB_itempar_m2;
run;


*Fit the Multidimensional IRT model with simple structure to the same data using
NLMIXED;

proc nlmixed data= DataB_sample_t method=gauss noad technique=quanew qpoints=7;
bounds a1_1-a1_42 >0;
parms a1_1-a1_42=1 b1-b42=0 d14_1=0 d28_1=0 d42_1=0 cov_12=0.5 cov_13=0.5
cov_23=0.5;
beta = b1   *i1   +
b2    *i2   +
b3    *i3   +
b4    *i4   +
b5    *i5   +
b6    *i6   +
b7    *i7   +
b8    *i8   +
b9    *i9   +
b10   *i10  +
b11   *i11  +
b12   *i12  +
b13   *i13  +
b14   *i14  +
b15   *i15  +
b16   *i16  +
b17   *i17  +
b18   *i18  +
b19   *i19  +
b20   *i20  +
b21   *i21  +
b22   *i22  +
b23   *i23  +
b24   *i24  +
b25   *i25  +
b26   *i26  +
b27   *i27  +
b28   *i28  +
b29   *i29  +
b30   *i30  +
b31   *i31  +
b32   *i32  +
b33   *i33  +
b34   *i34  +
b35   *i35  +
```

```
b36   *i36  +
b37   *i37  +
b38   *i38  +
b39   *i39  +
b40   *i40  +
b41   *i41    +
b42   *i42
;
slope1=
a1_1  *i1   +
a1_2  *i2   +
a1_3  *i3   +
a1_4  *i4   +
a1_5  *i5   +
a1_6  *i6   +
a1_7  *i7   +
a1_8  *i8   +
a1_9  *i9   +
a1_10 *i10  +
a1_11 *i11  +
a1_12 *i12  +
a1_13 *i13  +
a1_14 *i14  +
a1_15 *i15  +
a1_16 *i16  +
a1_17 *i17  +
a1_18 *i18  +
a1_19 *i19  +
a1_20 *i20  +
a1_21 *i21  +
a1_22 *i22  +
a1_23 *i23  +
a1_24 *i24  +
a1_25 *i25  +
a1_26 *i26  +
a1_27 *i27  +
a1_28 *i28  +
a1_29 *i29  +
a1_30 *i30  +
a1_31 *i31  +
a1_32 *i32  +
a1_33 *i33  +
a1_34 *i34  +
a1_35 *i35  +
a1_36 *i36  +
a1_37 *i37  +
a1_38 *i38  +
a1_39 *i39  +
a1_40 *i40  +
a1_41 *i41   +
a1_42 *i42
;


d1= d14_1*i14 + d28_1*i28 + d42_1*i42;
eta1= (slope1*theta1*p1 + slope1*theta2*p2 + slope1*theta3*p3- beta)*(i1      +
i2    +
```

```
i3    +
i4    +
i5    +
i6    +
i7    +
i8    +
i9    +
i10   +
i11   +
i12   +
i13   +
i15   +
i16   +
i17   +
i18   +
i19   +
i20   +
i21   +
i22   +
i23   +
i24   +
i25   +
i26   +
i27   +
i29   +
i30   +
i31   +
i32   +
i33   +
i34   +
i35   +
i36   +
i37   +
i38   +
i39   +
i40   +
i41)
+(slope1*theta1*p1+slope1*theta2*p2+slope1*theta3*P3 - beta-d1)*(i14+
i28+i42);
num1 = exp(eta1);
num2 = exp(2*eta1+2*d1)*(i14+i28+i42);
denom=1+num1+num2;
if (score1=0) then prob=1/denom;
else if (score1=1) then prob=num1/denom;
else if (score1=2) then prob=num2/denom;
if(prob>1E-8) then ll=log(prob);
else ll=-1E100;
model score1 ~ general(ll);
random theta1 theta2 theta3 ~ Normal ([0,0,0],[1,Cov_12,1,Cov_13,Cov_23, 1])
subject= person;

ods output ParameterEstimates= DataB_itempar_m3;
run;
```