Psicológica (2016), 37, 85-104.

Modeling Local Item Dependence in Cloze and Reading Comprehension Test Items Using Testlet Response Theory

Purya Baghaei[1,*], and Hamdollah Ravand[2]

[1]English Department, Islamic Azad University, Mashhad Branch, Mashhad, Iran

[2]Vali-e-Asr University of Rafsanjan, Iran, and University of Jiroft, Iran

In this study the magnitudes of local dependence generated by cloze test items and reading comprehension items were compared and their impact on parameter estimates and test precision was investigated. An advanced English as a foreign language reading comprehension test containing three reading passages and a cloze test was analyzed with a two-parameter logistic testlet response model and a two-parameter logistic item response model. Results showed that the cloze test produced substantially higher magnitudes of local dependence than reading items, albeit the levels of local dependency produced by reading items was not ignorable. Further analyses demonstrated that while even substantial magnitudes of testlet effect does not impact parameter estimates it does influence test reliability and information. Implications of the research for foreign language proficiency testing, where testlets are regularly used, are discussed.

Testlets are sets of items grouped together around the same stimuli such as shared reading passages, scenarios, figures, or tables. Testlets have been lauded in educational testing on the following grounds: (a) they save testing time as it is more efficient both for test developers and test takers to have a number of items following a common stimulus than to have just a single item, (b) unlike atomistic decontextualized items such as multiple-choice items, testlets, which are a combination of linked items, may

[*] Corresponding author: Islamic Azad University, Ostad Yusofi St., 91871-Mashhad, Iran. Phone: +98 51 36635064; Fax: +98 51 36634763; E-mail: pbaghaei@mshdiau.ac.ir

increase authenticity of test tasks by providing more context, (c) testlets provide solutions to some of the problems associated with adaptive tests. In adaptive tests, where examinees take dissimilar sets of items, context effects due to item location, cross information, or unbalanced content may introduce construct-irrelevant variance into the assessment. Testlets can diminish these contextual effects by forming fixed item-content units (Wainer, Bradlow, & Wang, 2007).

Despite their appealing features, testlets may introduce additional sources of construct-irrelevant variance. Items grouped under the same testlet might correlate with each other over and above the influence of the latent trait. The interrelatedness is likely to lead to a problem known as *local item dependence* (LID) in educational testing. A critical assumption in all standard statistical models in general and educational testing in particular is independence of observations. Items are said to be locally dependent if a person's response to an item is dependent on his response to another item. Local independence assumption is obtained when persons' responses to test items are affected solely by the trait intended to be measured by the items. When the contribution of the intended latent trait is removed, correlations between the items should be zero, unless there is a secondary dimension affecting responses. This subsidiary dimension might arise due to person-related characteristics such as differences in motivation or attention, differences in background knowledge, and ambiguities in the information provided in the input (Yen, 1993; Yen & Fitzpatrick, 2006). Using standard item response theory (IRT) models when LID is present may lead to problems such as biased item difficulty and discrimination parameters, overestimation of the precision of person ability estimates, overestimation of test reliability and test information, and underestimation of the standard errors of parameter estimates (Wainer et al., 2007; Yen & Fitzpatrick, 2006). Ignoring LID might lead to sever problems in judging psychometric qualities of tests which might in turn result in serious consequences regarding test score interpretation and use. In computer adaptive tests, for example, overestimation of the precision of person parameters might lead to premature termination of the test, where precision of ability estimate is the termination criterion (Wainer & Wang, 2000). Also in classical test theory LID can lead to overestimation of reliability which is the result of high intercorrelations among items in the same testlet over and above the construct of interest. Zhang (2010) showed that ignoring testlet effect can also diminish classification accuracy of examinees.

LID has been addressed in one of the following ways in the litrature: (1) Score-based polytomous item response theory models such as the graded response model (Samejima, 1969), polytomous logistic regression

(Zumbo, 1999), partial credit model (Masters, 1982), and rating scale model (Andrich, 1978) have been fitted to testlet data (Baghaei, 2010). In these models each testlet with *m* questions is treated as a super item with the total score ranging from 0 to *m*, (2) Item-based testlet response theory models (TRT) such as the 2-PL TRT (Bradlow, Wainer, & Wang, 1999), 3-PL TRT (Wainer, Bradlow, & Du, 2000), and the Rasch testlet model or 1-PL TRT (Wang & Wilson, 2005) have been employed, and (3) Item-based multilevel testlet response thoery models such as the three-level testlet response theory model (Jiao, Wang, & Kamata, 2005) or the two-level cross-classified testlet response theory model (Beretvas & Walker, 2012; Ravand, 2015) have been employed.

Score-based approaches to LID are limited in that: (1) They would lead to loss of information since they do not take into account the exact response patterns of test takers to individual items within a testlet, that is, the difference in the response patterns of the examinees with the same sum scores is not known (Wainer, et al., 2007) and (2) the model only works when LID magnitude is moderate and there are many independent items (Wainer, 1995).

To account for LID without loss of information, Bradlow, Wainer, and Wang (1999) advanced a TRT model which is an extension of the 2 PL IRT model (Birnbaum, 1968). The 2 PL TRT is multidimensional IRT model (Reckase, 2009; see also Baghaei, 2012) which includes a random effect parameter, γ, to account for the interdependencies of the items within the same testlet. According to this model, the probabiltiy of a correct answer to an item i nested in testlet $d(i)$for a person n with ability θ_n is expressed as:

$$p(x_i = 1 \mid \theta_n) = \frac{\exp[\alpha_i(\theta_n - \mathrm{b}_i - \gamma_{nd(\mathrm{i})})]}{1 + \exp[\alpha_i(\theta_n - \mathrm{b}_i - \gamma_{nd(\mathrm{i})})]}, \tag{1}$$

where $\alpha_i$and $\mathrm{b}_i$are the item discriminationation and difficulty parameters, respectively, and $\gamma_{nd(\mathrm{i})}$is the testlet effect parameter for person$n$ on testlet $d(i)$.

The distinctive component of the model, as compared to the standard 2-PL model is the introduction of the random effects parameter γ, which represents the local dependency within each testlet $d(i)$. TRT yields two person ability parameters: a general abiltiy $\theta_n$and a testlet specific ability $\gamma_{nd(\mathrm{i})}$. The testlet specific parameter is caused by person characteristics such as background knowledge, passage dependence, motivation, etc. $\gamma_{nd(\mathrm{i})}$is common (i.e. fixed) across items and random (i.e., varying) across persons (Wainer et al., 2007; Wang & Wilson, 2005). Introduction of the random effects parameter makes TRT a special case of multidimensional IRT models, i.e., a bifactor model, where each item simultaneously loads on two factors (dimensions): an overall ability dimension and a testlet-specific dimension.

When the $\gamma = 0$, i.e. when there is no testlet effect, the assumption of local depnedence holds and the model reduces to a standard 2-PL model. The higher the variance of γ the greater the LID. It's worthy of note that Equation 1 reduces to the Rasch testlet model if the discrimination paremeter α_i is the same for all the items.

TRT models have been applied extensively to model LID. Wainer and Wang (2000) applied the 3-PL TRT model to analyze LID in the reading and listening comprehension sections of the Test of English as a Foreign Language (TOEFL). They found that LID due to testlet effect did not affect difficulty estimates but resulted in biased discriminationa and guessing parameter estimates. They also compared the 3-PL TRT results with those obtained from a standard IRT model. They found that when LID was ignored, there was an overestimation in test information by 15% for some ability levels. In another application of the 3-PL TRT, Chang and Wang (2010) explored testlet effect in the Programme for International Student Assessment. In line with Wainer and Wang's (2000) study they found negligible effect of LID on item difficulty estimates. However, they found item discrimination and the precision of examinee ability meaures were overestimated. Zhang (2010) studied the effect of ignoring LID on examinee classification accuracy. He found that standard errors of abiltiy estimates obtained under the 3-PL TRT were sizably higher than those based on the standard IRT model, erroniously implying higher measurement precision in the estimates. Along the same lines, in the context of mastery classification for criterion referenced tests Baghaei (2007) demonstrtaed that ignoring LID can lead to erronious decisions especially near the cut-score.

In two more recent studies Eckes (2014) and Eckes and Baghaei (2015) employed the 2-PL TRT model to explore testlet effects. Eckes (2014) studied testlet effect in the listening section of Test of German as a Foreign Language (TestDaf). He found that ignoring LID led to overestimation of relaibility and underestimation of standard errors of ability estimates. He also found that item discrimination and difficulty estimates were not severly affected. Eckes and Baghaei (2015) examined local dependence in the items of a C-test. They found that testlet effects were very small hence parameters obtained under the 2-PL TRT and those obtained from the standard IRT models were highly comparable. They also found that when LID was ignored the C-test reliability was overestimated.

Cloze test and reading comprehension

A classical cloze test consists of a single longer passage in which every nth word is deleted where the test takers have to supply the missing words (Oller, 1979). Researchers are divided on the issue of what cloze tests measure. The long-running argument has largely concerned whether cloze tests measure sentence-level grammatical knowledge or global text comprehension. Some studies have argued that cloze tests are appropriate for measuring reading comprehension ability (Chihara, Oller, Weaver, & Chavez-Oller, 1977; Bachman, 1985; Jonz, 1990; McKenna and Layton, 1990; Chavez-Oller, Chihara, Weaver, & Oller, 1985). They have shown that cloze tests are sensitive to text-level constraints. On the other hand, some other studies have concluded that cloze tests are measures of local syntactic constraints (e.g., Alderson, 1979; 1980; Kibby, 1980; Shanahan et al., 1982; Markman, 1985). However, there is “some consensus among researchers that not all deletions in a given cloze passage measure exactly the same abilities” (Bachman, 1985, p.535). Bachman (1985) concludes that a possible source of inconsistency among the results of studies on construct validity of cloze tests is that these studies have not distinguished between the cloze types. The majority of the studies have created the cloze tests through the fixed-ratio deletion procedure (every nth word deleted). Bachman prepared two versions of cloze from the same passage: fixed-ratio and rational cloze. He classified the knowledge type required to restore the missing words in the rational cloze as: (1) within clause, (2) within sentence, (3) within text, and (4) extra-textual. He concluded that the majority of the words deleted in the every-nth-words-deleted procedure (i.e., fixed-ratio) were of the clause-level (Type 1) or the extra-textual (Type 4) and there were few gaps requiring within-sentence and within-text context. The results of his study also indicated that through a rational deletion method, test developers can include deletions of the Types 2 and 3 which can be restored using textual understanding. In line with the finding of Bachman, Alderson (2000) reserved the term cloze for the fixed-ratio procedure and “gap-filling” for the conventionally called rational deletion cloze tests. He argues that gap-filling tests can be used as reading comprehension tests. Yamashita (2003) investigated construct validity of a gap-filling test using verbal protocols of the test-takers’ response processes categorized according to the classification framework developed by

Bachman (1985). The results of his study showed that readers at different ability levels used text-level processing more frequently. He concluded that gap-filling tests can be used to measure higher-level reading comprehension ability.

Along the same lines, Greene (2001) found out that rational deletion cloze tests and true-false reading comprehension tests produce the same mean and dispersion among college level students. Greene (2001) argued that a cloze test whose items have been carefully selected to tap into test takers' inference making ability can "measure a reader's macroprocessing of theoretical text" (p. 92).

Bormuth (1969) studied the factorial validity of cloze test by administering nine cloze tests and seven multiple-choice reading comprehension passages to a sample of grade four, five, and six students. Exploratory factor analysis yielded one factor with an eigenvalue greater than one which accounted for 77% of the variance. All cloze tests and reading comprehension passages loaded on this factor which was named reading comprehension ability factor. Further correlational and factor analytic evidence for the validity of cloze as a measure of reading comprehension has been provided by Rankin (1959), Bormuth (1967, 1968a, 1968b), Rye (1982), Harrison, (1980), and Klare, (1984).

Purpose of the study

One key issue in cloze tests which poses problems for the analysis and scaling of items and persons is the interdependency of cloze test items. Dependency among cloze items is a violation of the local item independence assumption addressed above. This restricts analysis of cloze test items with IRT models and the estimation of its reliability with internal consistency methods such as Cronbach's alpha (Bachman, 1990; Farhady, 1983).

The same problem is encountered for the analysis of reading comprehension items as such items are usually clustered around a passage which makes them a testlet and hence may lead to the violation of local independence. In language testing literature it is commonly stated that due to the interconnected structure of cloze test items internal consistency reliability estimates such as Cronbach's alpha are not appropriate for estimating reliability and test-retest and parallel-forms methods are suggested instead (Bachman, 1990; Farhady, 1983). This admonition is also given for C-Test which is a special form of cloze test. In a C-Test instead of deleting entire words the second half of every other word is deleted and examinees have to restore the missing letters (Grotjahn1987; Klein-Braley, 1997). As a C-Test battery is always composed of four to six short texts correct answers on each text is aggregated and each passage is entered into the analysis as a polytomous item (Baghaei, 2011; Raatz &Klein-Braley, 2002; Sigott, 2004). Obviously this is done to circumvent the local dependence which seems to be present in C-Test.

In this study we aim to address the issue of LID in cloze tests and reading comprehension tests by comparing the magnitude of LID each generates and its effect on parameter estimates and test precision. To address these issues the following research questions were formulated:

1. To what extent reading comprehension test items and the cloze test items introduce local item dependency (LID)? Which test type, whether cloze or reading, is more affected by local dependency? In other words, which kind of test is a more intense source of LID? The general impression is that due to the interconnected structure of items in cloze tests they induce more dependency (Bachman, 1990; Farhady, 1983) than reading comprehension items which are at least structurally independent of each other. It is interesting to put this assumption into an empirical test and compare the two test types in terms of generating LID.

2. How do person and item parameters obtained under the standard 2-PL IRT model, where LID is ignored, compare with those obtained under the 2-PL TRT model where LID is accounted for? Answer to this question helps gauge practical consequences of local dependence for measurement in the contexts where testlets are regularly used.

METHOD

Instrument and Participants. Participants of the study were two random subsamples of the Iranian National University Entrance Examination (INUEE) candidates (N_1=5412, 71.3 % females and 26.8 % males and N_2=5374, 69.3 % females and 28.9 % males) who applied for the English master programs in state universities in 2012. INUEE, as the sole selection criterion, is a high-stakes test that screens the applicants into English Studies programs at M.A. level in state-run universities in Iran. The INUEE is administered once a year in February. The participants were Iranian nationals and the majority of them held a B.A. degree in English Studies (88 %).

The INUEE measures general English proficiency at an advanced level and content knowledge of the applicants in teaching methodology, principles of language testing, and linguistics. The general English section consists of 10 grammar items, 20 vocabulary items, 20 reading comprehension items, and 10 cloze test items. Cloze is deemed to be a test of general language proficiency in a foreign language (Oller, 1972, 1979) and a test of reading comprehension by other researchers (Bormuth, 1969; Greene, 2001; Rye, 1982). Inspection of the cloze test analysed in the present study clearly revealed that the deletion method is not fixed-ratio deletion as the distances among the gaps were not equal. Most of the deletions were cohesive devices and key content words which required text-level understanding. Therefore in can be argued that the cloze test is a rational-deletion cloze and, therefore, a test of reading comprehension. Nevertheless, there is no information in the test documentation on the rationale behind the deletions.

All the questions were multiple-choice and test takers had to complete the items in 60 minutes. For the purpose of the present study only the reading comprehension section and the cloze test were used. The reading comprehension section was composed of three passages on academic topics with seven, six and seven items following each passage, respectively. The cloze test was 4-option multiple choice with 10 items (gaps).

Data analysis. Two IRT models were separately fitted to the selected samples of the study: (1) a standard 2-PL IRT model (Birnbaum, 1968) where it is assumed that no LID exists and (2) a 2-PL TRT model (Bradlow & Wainer, & Wang, 1999), where LID is systematically modeled and conditioned out[1]. SCORIGHT computer programme (Version 3.0; Wang, Bradlow, & Wainer, 2005) was used for the analyses. To estimate the model parameters Bayesian estimation techniques are implemented in SCORIGHT. Bayesian methods incorporate prior information about model parameters to facilitate estimation (Fox, 2010; Gelman, Carlin, Stern, & Rubin, 2003). In SCORIGHT, inferences for unknown parameters are obtained by drawing samples from their posterior distributions using Markov Chain Monte Carlo (MCMC) techniques (Wang, et al., 2005). The 2-PL IRT and 2-PL TRT models were fitted to the data using Markov chain Monte Carlo (MCMC) methods. Five chains were run. For each chain, after a burn-in period of 4000 iterations, as is advised to be sufficient by Wang, et al. (2005), the next 1000 iterations were used for inferences, where every tenth draw was retained to reduce the high autocorrelation. The MCMC sampler converged properly as the potential scale reduction factors for the prior and hyperprior parameters were all very close to 1.0.

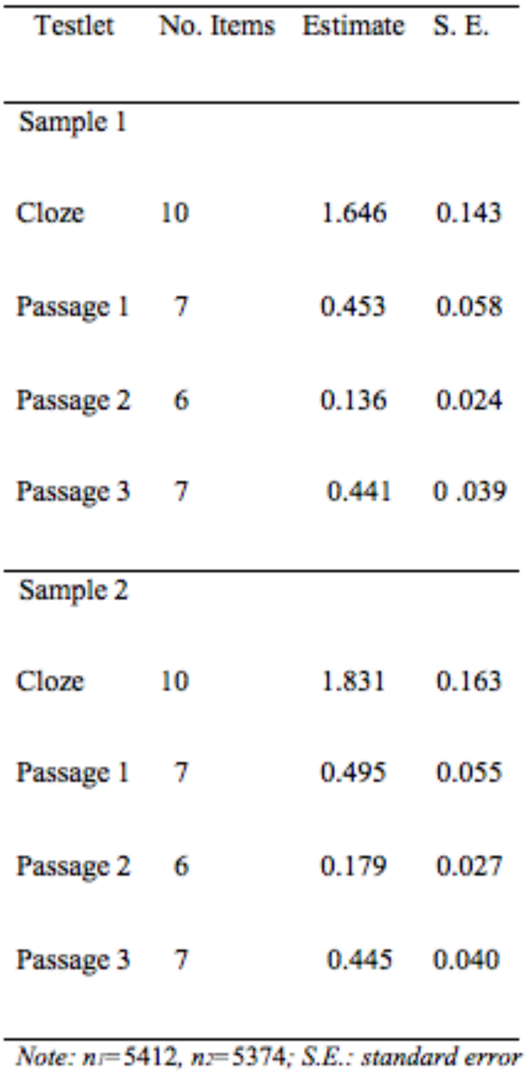
[1] Since the test is multiple-choice we first fitted the 3-PL IRT and 3-PL TRT models to account for guessing too. But the model did not converge although we ran 10 chains with 35000 iterations in each chain.

RESULTS

Testlet effects

Table 1 presents the magnitudes of γ or testlet effects and their associated standard errors for each testlet in the two samples. As the table shows, cloze test generates the highest level of local dependency. This finding agrees with our expectations considering the structure of cloze test items.

Table 1. Testlet statistics in samples 1 and 2.



Testlet	No. Items	Estimate	S. E.
Sample 1			
Cloze	10	1.646	0.143
Passage 1	7	0.453	0.058
Passage 2	6	0.136	0.024
Passage 3	7	0.441	0 .039
Sample 2			
Cloze	10	1.831	0.163
Passage 1	7	0.495	0.055
Passage 2	6	0.179	0.027
Passage 3	7	0.445	0.040

Note: n_1=5412, n_2=5374; S.E.: standard error

There are no accepted rule of thumb values for judging testlet effect parameters (Eckes & Baghaei, 2015). Simulation studies show that testlet effects smaller than .25 are negligible (Glas, Wainer, & Bradlow, 2000; Wang, Bradlow, & Wainer, 2002; Wang & Wilson, 2005; Zhang, 2010). Note that TRT is a bifactor model where items nested within a testlet are modeled to load on a specific testlet dimension (a group factor) while simultaneously loading on a general ability dimension. Testlet effect parameters, γ, are in fact the variances of these specific factors. To judge the magnitude of the local dependence generated by a testlet, γ is compared with the variance of the general ability dimension. The higher the variance of testlet specific dimensions compared to the variance of the general ability dimension, the more local dependence the testlet has generated (Baghaei & Aryadoust, 2015). In SCORIGHT for model identification the variance of the ability distribution is set to one with mean of zero. Therefore, testlet effects are compared with one. In Sample 1 the cloze test has a testlet effect equal to 1.646 which is substantially higher than the variance of the general ability dimension. Reading passages 1 and 3 have testltet effect estimates almost half the variance of the general ability dimension, which are not negligible. Reading passage 2 exhibits a benign magnitude of local dependency. Testlet effects are slightly higher in sample 2 but the same pattern is observed.

Item parameters

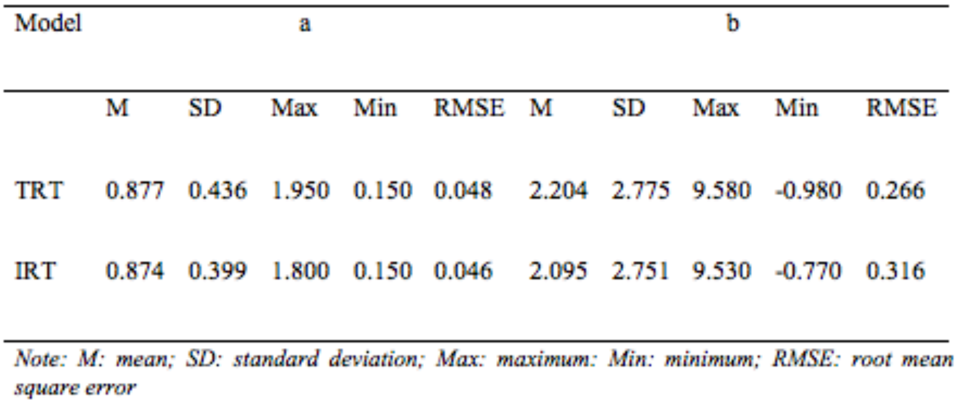
In standard IRT it is assumed that no local dependency among items exists while in TRT local dependency is factored out systematically by adding a random effect parameter γ to the IRT item response function. In this section item parameters across the two models, i.e., TRT which accounts for LID and IRT where LID is ignored are compared.

Table 2 shows the descriptive statistics for item parameters in the two models in sample 1. The root mean-square measurement error (RMSE) is the square root of the average of the parameter error variances which is an index of precision of estimation (Linacre, 2012).

In Sample 1 discrimination parameters across the two models correlated at .989. The difference between *a* parameters estimated by each model for each item was computed. The absolute values of these differences ranged from 0.00 to .210 with a mean of .052 and a standard deviation of .049. The mean of squared absolute differences turned out to be .005 with a standard deviation of .008. The root mean square deviation was .070.

Difficulty parameters across the two models correlated at .997. The difference between *b* parameters estimated by each model for each item was computed too. The absolute values of these differences ranged from 0.00 to .640 with a mean of .152 and a standard deviation of .164. The mean of squared differences turned out to be .049 with a standard deviation of .095. The root mean square deviation was .221. RMSE's show that while discrimination parameter is estimated with the same precision across the models the difficulty parameter is estimated with a higher precision under TRT.

Table 2. Summary statistics for discrimination and difficulty parameters under TRT and IRT.



Model	a					b				
	M	SD	Max	Min	RMSE	M	SD	Max	Min	RMSE
TRT	0.877	0.436	1.950	0.150	0.048	2.204	2.775	9.580	-0.980	0.266
IRT	0.874	0.399	1.800	0.150	0.046	2.095	2.751	9.530	-0.770	0.316

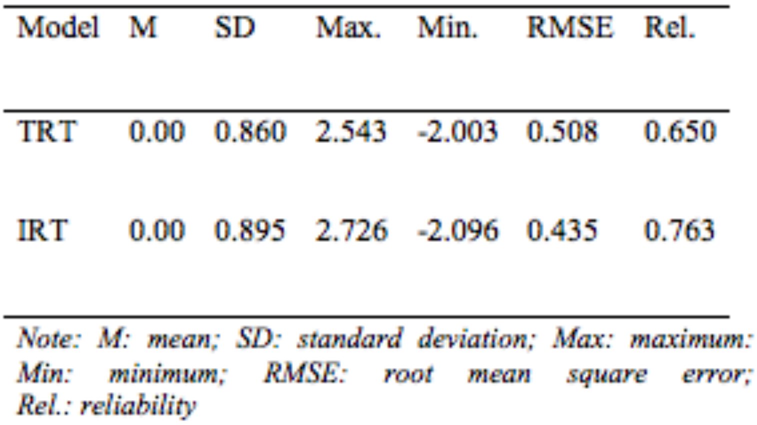
Note: M: mean; SD: standard deviation; Max: maximum: Min: minimum; RMSE: root mean square error

Results in sample 2 were highly comparable with those in sample 1 as discrimination parameters across the two models correlated at .990 and difficulty parameters correlated at .997. The other indices were highly similar to those obtained in Sample 1, so they are not reported here. The discrimination parameters estimated across the two independent samples correlated at .866 under IRT and .877 under TRT. The difficulty parameters estimated in the two independent samples correlated at .803 under IRT and .812 under TRT. This shows that while parameter estimates remain relatively stable across independent non-overlapping samples the discrimination parameter has remained more invariant. IRT and TRT performed equally well as far as the stability of parameters estimation across different populations is concerned, with TRT slightly performing better.

Person parameters

In Sample 1 person parameters across the two models correlated at .994 which indicates considerable correspondence in the two models. The difference between theta parameters estimated by each model for each person was computed. The absolute value of these differences ranged from 0.00 to .380 with a mean of .079 and a standard deviation of .059. The mean of squared differences turned out to be .009 with a standard deviation of .013. The root mean square deviation was .094. Table 3 presents the summary statistics for the theta parameters in the two models along with root mean square errors and reliabilities.

Table 3. Summary statistics for person ability estimates under TRT and IRT.



Model	M	SD	Max.	Min.	RMSE	Rel.
TRT	0.00	0.860	2.543	-2.003	0.508	0.650
IRT	0.00	0.895	2.726	-2.096	0.435	0.763

Note: M: mean; SD: standard deviation; Max: maximum: Min: minimum; RMSE: root mean square error; Rel.: reliability

Note that RMSE is the square root of the mean of the theta parameters' squared standard errors. The reported reliabilities are Bayesian reliabilities (Wainer, et al., 2007) which are computed by dividing the variance of the expected a posteriori theta parameters by the same value plus the mean of squared person parameters standard errors. The mean of the ability distribution is set to zero for model identification. Table 4 shows that both models have produced the same amount of dispersion with IRT yielding a slightly wider distribution. The major difference in the two models is the smaller error of estimation and higher reliability of IRT estimates compared to the TRT model.

In sample 2 person parameters across the two models correlated at .996 which indicates considerable correspondence in the two models. The rest of the agreement statistics were highly comparable to those obtained for Sample 1. As was observed in Sample 1, the striking difference was in the reliability of the estimates. The reliability for TRT person parameters was .746 and for IRT was .814.

DISCUSSION

The assessment of reading comprehension almost always entails testlets. Reading comprehension tests make use of passages which are followed by several questions. This may be supplemented by a passage with missing words or letters to be filled in. Although using testlets in educational measurement is economic, efficient, and valid (Wainer, et al., 2007) they pose problems for data analysis: Testlets violate the conditional independence assumption of IRT models and can lead to biased parameter estimates and spuriously low standard errors (Baghaei, 2010; Baghaei & Aryadoust, 2015; Wainer & Wang, 2000). The purpose of this investigation was (1) to examine the extent to which cloze test items and reading comprehension items generate local item dependence and (2) to assess the impact of local item dependence on item and person parameter estimates and their precision.

Findings of the study indicated that, as expected, cloze test items generated substantially higher levels of local dependency than reading comprehension items. The LID magnitude in the cloze test was almost four times greater than the LID produced in reading passages. This is in line with Zhang (2010) who found a testlet effect magnitude of 1.43 for a cloze test with 20 items and testlet effect values of .58, .53, .35, and .59 for four reading comprehension passages (each with five items) in the Examination for the Certificate of Proficiency in English. However, the results are in contrast with those of Eckes and Baghaei (2015) and Schroeders, Robitzsch, & Schipolowski (2014) who found very small testlet effects for C-Test passages.

This discrepancy can be attributed to the differences in the structures of cloze test and C-Test. Frequent half-word deletions in C-Tests (in contrast to every 5th or 6th full-word deletions in cloze) probably prevents application of text level processes by examinees. In other words, C-Test taking is more a local gap filling task without resort to higher-order text level processes. Another possible reason for the contrasting results between cloze and C-Test might be the language-specific characteristics and types of the texts used to construct the cloze tests and the C-Tests used in these studies. Another reason for the different magnitudes of LID generated by C-Test and cloze test might be the text length. In cloze tests usually longer passages are used but in C-Tests several independent short passages are employed. Obviously short passages do not allow for text-level processes to be activated by the examinees.

The reason for higher magnitude of LID in the cloze testlet compared to the reading comprehension testlets could be due to the fact that LID in cloze has two sources. Marais and Andrich (2008) distinguished between two types of LID: *trait dependence* and *response dependence*. Trait dependence (referred to as *multidimensionality* in the literature) occurs when a secondary trait or dimension is being measured by the test. Response dependence or *item chaining effect* (Wang & Wilson, 2005) occurs when answer to an item affects how subsequent items are answered. Higher LID for cloze tests are expected since they are more likely to be affected by both trait and response dependence.

Another reason for the higher magnitude of dependency in the cloze test could be the method specific features of the cloze that may require certain abilities over and above language proficiency. As explained earlier, TRT is a bifactor model in which all items load on a general ability dimension and at the same time each item loads on a testlet specific dimension. One shortcoming of TRT is that unique variances in testlets are lumped together and assumed to be the variance due to LID only. While the unique variance could be a combination of many other factors such as testlet specific knowledge or test method variance (Baghaei & Aryadoust, 2015).

Unless we can argue that the cloze and reading passages are measures of the same construct, comparison of testlet variances across these two test types is not justified. The higher magnitude of dependency in the cloze test might simply reflect cloze specific abilities not shared with reading comprehension and not construct-irrelevant variance due to item dependency. As it was argued before, the literature on the construct validity of rational cloze tests has demonstrated that they measure higher-level reading comprehension ability.

There is no way to disentangle these variances unless we have multiple cloze testlets as well as multiple reading testlets. This can be addressed with at least two cloze tests and two reading comprehension passages. A bifactor model is run and all items (cloze and reading) are forced to load on a reading factor while the cloze tests load on a cloze factor as well. Furthermore, each testlet loads on a testlet specific factor. This way we can separate testlet variance from cloze specific variance in cloze passages.

Despite observing large to huge testlet effects in this study a great agreement was observed in person and item parameter estimates across the two models as was shown by extremely high correlations between the parameters estimated by the two models. On the surface, the level of local dependence did not affect the estimates of parameters of interest in IRT where local dependence was ignored. But the observed differences in the estimates across the models for some items and persons were as high as 0.21, .64, and 0.38 for discrimination, difficulty, and person ability parameters, respectively. Such differences could have adverse consequences when high-stakes decision making is involved. Furthermore, the two models performed equally well in in terms of stability of parameter estimation across different populations.

Perhaps the most important ramification when the inappropriate model is employed is the reliability of estimates. It is generally argued that local dependency leads to biased parameter estimates including inflated item discrimination estimates, overestimation of precision of person ability estimates, and overestimation of test reliability and test information (Baghaei, 2010; Zhang, 2010). The results of the present study showed that despite observing very high magnitudes of testlet effect item and person parameters across the two models were closely comparable. The only substantial difference observed between the two models was in reliability and the precision of ability estimates. IRT overestimated test reliability and the precision of person ability parameters. This can lead to serious problems when computer adaptive testing is used, where the criterion for test termination is the standard error of person estimates, i.e., it leads to premature test termination (Wainer & Wang, 2000). Ip (2010) showed that when LID exists information functions are wrong and testlet effect, which is present in language assessments, results in overestimation of classification accuracy due to the underestimated measurement error (Zhang, 2010).

The TRT approach employed in this study, models testlet effect as random. However, testlet effects can also be assumed as *fixed* (Beretvas & Walker, 2012). In the fixed-effects approach testlet effect is assumed to be constant over persons. In other words, LID is an item characteristic. On the other hand, in the random-effects approach testlet effects are assumed to be varying (random) over persons; testlet effects are different for people with different learning experiences, background knowledge, or levels of interest. According to Wang and Wilson (2005), the random-effects approach is more appropriate where trait dependence is present, whereas the fixed-effects approach is more suitable where item-chaining effect or response dependence is suspected. With tests such as cloze where both types of dependencies are suspected, an approach which takes into account both fixed and random effects would be more appropriate. The two-level testlet response model (MMMT-2) proposed by Beretvas and Walker (2012) models both fixed and random testlet effects. The random testlet effect in MMMT-2 corresponds with a secondary dimension that the testlet might be measuring. In other words, the random testlet effect is the effect of the testlet on the difficulty of the items within the testlet, which is due to the secondary dimension targeted by the testlet. However, the fixed testlet effect can be interpreted as the direct contribution of the testlet to the difficulty of an item on the primary dimension being measured by the item. In other words, the fixed testlet effect represents the effect of testlet on the diffuculty of the items, which is due to the primary dimension intended to be assessed by the testlet.

Limitations and Suggestions for Further Research

The current study compared the magnitude of local item dependence generated by cloze and reading comprehension items and its impact on parameters estimates and their precision in an advanced test of English as a foreign language. Findings showed that while even substantial magnitudes of testlet effect does not impact parameter estimates it does influence the test reliability and information.

The generalizability of the present study is limited in that there were relatively few testlets, especially for the cloze test, and too many test takers. To better compare testlet effect in cloze and reading comprehension tests, further research can include more cloze and reading testlets. Generalization of the findings of the present study might be further limited by the difficulty of the items studied. It is suggested future studies compare testlet effect in testlets with easier items. Since a prerequisite in comparing LID across cloze and reading comprehension tests is that they measure the same latent trait, it is suggested that future studies develop rational cloze tests with gaps according to the framework suggested by Bachman (1985) with gaps of Type 2 and 3 (i.e., within sentence and within text) and then compare LID across reading and cloze. As noted before another strategy is to have multiple cloze tests and multiple reading testlets and use a bifactor model to disentangle the testlet effect from the cloze specific variance in cloze tests.

REFERENCES

Alderson, J. C. (1979). The cloze procedure as a measure of proficiency in English as a foreign language. *TESOL Quarterly*, *13*, 219-227. DOI: 10.2307/3586211.

Alderson, J. C. (1980). A process approach to reading at the University of Mexico. *ELT Documents: Special Issue*. *Projects in Materials Design*, 134-143.

Alderson, J. (2000). *Assessing reading*. New York: Cambridge University Press

Bachman, L. F. (1985). Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly*, *19*, 535-556. DOI: 10.2307/3586277.

Bachman, L.F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Baghaei, P. & Aryadoust, V. (2015). Modeling local item dependence due to common test format with a multidimensional Rasch model. *International Journal of Testing, 15*, 71-87. DOI: 10.1080/15305058.2014.941108.

Baghaei, P. (2012). The application of multidimensional Rasch models in large scale assessment and validation: An empirical example. *Electronic Journal of Research in Educational Psychology, 10*, 233-252.

Baghaei, P. (2011). Optimal number of gaps in C-Test passages. *International Education Studies*, *4*, 166-171. DOI: 10.5539/ies.v4n1p166.

Baghaei, P. (2010). A comparison of three polychotomous Rasch models for super-item analysis. *Psychological Test and Assessment Modeling, 52*, 313-322.

Baghaei, P. (2007). Local dependency and Rasch measures. *Rasch Measurement Transactions*. *21*, 1105-1106.

Beretvas, S. N., & Walker, C. M. (2012). Distinguishing differential testlet functioning from differential bundle functioning using the multilevel measurement model. *Educational and Psychological Measurement, 72*, 200-223. DOI: 10.1177/0013164411412768.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395-479). Reading, MA: Addison-Wesley.

Bormuth, J. R. (1967). Comparable cloze and multiple-choice comprehension test scores. *Journal of Reading*, *10*, 291-299.

Bormuth, J. R. (1968a). Cloze test readability: Criterion reference scores. *Journal of Educational Measurement*, *5*, 189-196. DOI: 10.1111/j.1745-3984.1968.tb00625.x.

Bormuth, J. R. (1968b). The cloze readability procedure. *Elementary English*, *45*, 429-436.

Bormuth, J. R. (1969). Factor validity of cloze tests as measures of reading comprehension ability. *Reading Research Quarterly*, *4*, 358-365. DOI: 10.2307/747144.

Bradlow, E. T., Wainer, H., & Wang, X., (1999). A Bayesian random effects model for testlets. *Psychometrika, 64*, 153-168. DOI: 10.1002/j.2333-8504.1998.tb01752.x.

Chang, Y., & Wang, J. (2010, July). *Examining testlet effects on the PIRLS 2006 assessment*. Paper presented at the 4th IEA International Research Conference, Gothenburg, Sweden.

Chavez-Oller, M.A., Chihara, T., Weaver, K.A., & Oller, J.W. Jr. (1985). When are cloze items sensitive to constraints across sentences? *Language Learning, 35*, 181-206. DOI: 10.1111/j.1467-1770.1985.tb01024.x.

Chihara, T., Oller , J. W. Jr .,Weaver, K. A. & Chavez -Oller, M. A. (1977). Are cloze items sensitive to constraints across sentences? *Language Learning, 27*, 63-70. DOI: 10.1111/j.1467-1770.1977.tb00292.x

Eckes, T., & Baghaei, P. (2015). Using testlet response theory to examine local dependence in C-Tests. *Applied Measurement in Education, 28,* 1-14. DOI: 10.1080/08957347.2014.1002919

Eckes, T. (2014). Examining testlet effects in the TestDaF listening section: A testlet response theory modeling approach. *Language Testing, 31*, 39-61. DOI: 10.1177/0265532213492969

Farhady, H. (1983). New directions for ESL proficiency testing. In J. W. Oller, Jr. (Ed.), *Issues in language testing research* (pp. 253–269). Rowley, MA: Newbury House.

Fox, J. P. (2010). *Bayesian item response modeling: Theory and applications*. New York: Springer. DOI: 10.1111/j.1751-5823.2011.00159_1.x.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd Ed.). Boca Raton, FL: Chapman & Hall/CRC.

Gilliland, J. (1972). *Readability*. London: University of London Press for the United Kingdom Reading Association.

Glas, C. A. W., Wainer, H., & Bradlow, E. T. (2000). MML and EAP estimation in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 271–288). Boston, MA: Kluwer-Nijhoff. Doi: 10.1007/0-306-47531-6_14.

Greene, B. (2001). Testing reading comprehension of theoretical discourse with cloze. *Journal of Research in Reading*, *24*, 82-98. DOI: 10.1111/1467-9817.00134.

Grotjahn, R. (1987). How to construct and evaluate a C-Test: a discussion of some problems and some statistical analyses. In R. Grotjahn, C. Klein-Braley, & D. K. Stevenson (Eds.) *Taking their measures: the validity and validation of language tests* (pp. 219-253). Bochum: Brockmeyer.

Harrison, C. (1980). *Readability in the classroom*. Cambridge: Cambridge University Press.

Ip, E. H. (2010). Interpretation of the three-parameter testlet response model and information function. *Applied Psychological Measurement, 34*, 467–482. DOI: 10.1177/0146621610364975

Jiao, H., Wang, S., & Kamata, A. (2005). Modeling local item dependence with the hierarchical generalized linear model. *Journal of Applied Measurement, 6*, 311-321.

Jonz, J. (1990). Another turn in the conversation: What does cloze measure? *TESOL Quarterly*, *24*, 61-83. DOI: 10.2307/3586852.

Kibby, M. W. (1980). Intersentential processes in reading comprehension. *Journal of Literacy Research*, *12*, 299-312. DOI: 10.1080/10862968009547383.

Klare, G. R. (1984). Readability. In P. D. Pearson (Ed.), *Handbook of reading research* (pp. 681-744). New York: Longman.

Klein-Braley, C. (1997). C-Tests in the context of reduced redundancy testing: An appraisal. *Language Testing*, *14*, 47-84. DOI: 10.1177/026553229701400104.

Linacre, J. M. (2012). *A user's guide to WINSTEPS-MINISTEP: Rasch-model computer programs*. Chicago, IL: Winsteps.com.

Markman, P.L. (1985). Rational deletion cloze and global comprehension in German. *Language Learning,* 35, 423–30. DOI: 10.1111/j.1467-1770.1985.tb01085.x.

McKenna, M. C., & Layton, K. (1990). Concurrent validity of cloze as a measure of intersentential comprehension. *Journal of Educational Psychology*, *82*, 372-377. DOI: 10.1037//0022-0663.82.2.372.

Oller, J.W. Jr. (1979). *Language tests at school*. London: Longman.

Oller, J.W. Jr. (1972). Scoring methods and difficulty levels for tests of proficiency in English as a second language. *Modern Language Journal, 56*, 151-158. DOI: 10.2307/324037.

Raatz, U. & Klein-Braley, C. (2002). Introduction to language testing and to C-Tests. In J. Coleman, R. Grotjahn, & U. Raatz, (Eds.) *University language testing and the C-Test* (pp. 75-91). Bochum: AKS-Verlag.

Rankin, E. F. (1959). The cloze procedure: its validity and utility. In *Eighth Yearbook of the National Reading Conference* (Vol. 8, pp. 131-144). Milwaukee: National Reading Conference.

Ravand, H. (2015). Assessing testlet effect, impact, differential testlet, and item functioning using cross-classified multilevel measurement modeling. *SAGE Open 5*, 1-9. DOI: 10.1177/2158244015585607.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York: Springer. DOI: 10.1007/978-0-387-89976-3.

Rye, J. (1982). *Cloze procedure and the teaching of reading*. London: Heinemann.

Samejima, F. (1969). *Estimation of ability using a response pattern of graded scores*. Psychometrika Monograph No. 17. Richmond, VA: Psychometric Society. DOI: 10.1007/bf02290599.

Schroeders, U., Robitzsch, A., & Schipolowski, S. (2014). A comparison of different psychometric approaches to modeling testlet structures: An example with C-Tests. *Journal of Educational Measurement, 51*, 400–418. DOI: 10.1111/jedm.12054.

Shanahan, T., Kamil, M. L., & Tobin, A. W. (1982). Cloze as a measure of intersentential comprehension. *Reading Research Quarterly*, *17*, 229-255. DOI: 10.2307/747485.

Sigott, G. (2004). *Towards identifying the C-Test construct*. Frankfurt/am: Peter Lang.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement*, *26*, 247–260. DOI: 10.1111/j.1745-3984.1989.tb00331.x.

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. DOI: 10.1207/s15324818ame0802_4.

Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An analog for the 3-PL useful in testlet-based 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. DOI: 10.1007/0-306-47531-6_13.

Wainer, H., Bradlow, E.T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge: Cambridge University Press.

Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement, 37*, 203-220. DOI: 10.1002/j.2333-8504.2001.tb01851.x.

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* (ETS Research Report RR 04-49). Princeton, NJ: Educational Testing Service. DOI: 10.1002/j.2333-8504.2004.tb01976.x.

Wang, W. C. & Wilson, M. (2005). The Rasch testlet model. .*Applied Psychological Measurement, 29*, 126–149. DOI: 10.1177/0146621604271053.

Wang, X., Bradlow, E. T., & Wainer, H. (2002). A general Bayesian model for testlets: Theory and applications. *Applied Psychological Measurement*, *26*, 109–128. DOI: 10.1177/0146621602026001007.

Yamashita, J. (2003). Processes of taking a gap-filling test: comparison of skilled and less skilled EFL readers. *Language Testing*, *20*, 267-293. DOI: 10.1191/0265532203lt257oa.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187-213. DOI: 10.1111/j.1745-3984.1993.tb00423.x.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111–153). Westport, CT: American Council on Education/Praeger.

Zhang, B. (2010). Assessing the accuracy and consistency of language proficiency classification under competing measurement models. *Language Testing, 27*, 119-140. DOI: 10.1177/0265532209347363.

Zumbo, B. D. (1999). *A Handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

(Manuscript received: 23 March 2015; accepted: 30 September 2015)