Using the Testlet Model to Mitigate Test Speededness Effects

James A. Wollack
Youngsuk Suh
Daniel M. Bolt

University of Wisconsin – Madison

April 12, 2007

Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

RUNNING HEAD: Mitigating Speededness Effects

Using the Testlet Model to Mitigate Test Speededness Effects

This paper studies the effectiveness of a three-parameter testlet mixed model (3PLt*) in accounting for local item dependence (LID) caused by test speededness. Data with varying amounts of speededness were simulated. Recovery of item and ability parameters was examined for the 3PLt*, a three-parameter mixture model for test speededness (M3PLM), and the three-parameter logistic model (3PLM). Results indicated that while the M3PLM recovered parameters slightly better than the 3PLt*, the 3PLt* presents a viable alternative. This is particularly true for trait estimation, because the legality of estimating ability for a subset of examinees using a different (less stringent) set of item difficulty estimates, as is done with the M3PLM, is questionable. In contrast, the 3PLM produced heavily biased estimates of ability, item difficulty and item discrimination for heavily speeded examinees, and failed to account for LID among end-of-test items.

Using the Testlet Model to Mitigate Test Speededness Effects

Tests consisting of items that violate the item response theory assumption of local item independence (LID) can cause serious problems for test developers. The inclusion of items with LID may result in spurious estimates of test reliability, item and test information, standard errors, item parameters, and equating coefficients (Lee, Kolen, Frisbie, & Ankenmann, 2001; Sireci, Thissen, & Wainer, 1991; Thissen, Steinberg, & Mooney, 1989; Wainer & Thissen, 1996; Yen, 1993). Depending on the nature of the cause of LID, examinees may suffer as well.

LID is commonly caused by having multiple items relate to a common stimulus, such as a reading passage (Thissen et al., 1989; Yen, 1993). One model which has proven very effective for accounting for this type of LID is the three-parameter testlet model (3PLt; Du, 1998; Bradlow, Wainer, & Wang, 1999; Wainer, Bradlow, & Du, 2000). The 3PLt explicitly models the systematic nuisance variation that commonly exists among items within a testlet by including into the model a random effects, testlet- and examinee-specific (parameter which is subtracted from the three-parameter logistic model (3PLM) item difficulty for examinee *j*. Du (1998), Wainer et al. (2000) and Li & Cohen (2003) found the 3PLt to work better than other available models for accounting for LID caused by testlets.

Another common cause of LID is test speededness (Yen, 1993). Speededness refers to testing situations in which some examinees do not have ample time to answer all questions. As a result, examinees may either hurry through, fail to complete, or randomly guess on items, usually at the end of the test. Unlike LID caused by testlets, speededness is usually an inadvertent source of LID in that the speed with which one responds is not an important part of the construct of interest. Examinees affected by test speededness typically show positive LID on items at the end of the test and receive ability estimates that underestimate their true levels. In addition,

speededness may cause certain items, particularly those administered late in the test, to have poorly estimated parameters (Douglas, Kim, Habing, & Gao, 1998; Oshima, 1994) making it difficult to hold together a score scale over time (Wollack, Cohen, & Wells, 2003).

In the past several years, a few models that explicitly model test speededness have been developed to improve the estimation of parameters for items at the end of the test. Bolt, Cohen, and Wollack (2002) developed a 2-class mixture item response model, with end-of-test items constrained to be harder in one class than in the other, to estimate item parameters separately for latent speeded and nonspeeded classes of examinees. Yamamoto & Everson (1997) developed a hybrid model which assumes that an item response model is appropriate throughout most of the test, but that items at the end of the test are answered randomly by some subset of examinees. Both the mixture and hybrid models have been shown to help improve the quality of item parameter estimates (Bolt, Mroch, & Kim, 2003), but the models suffer some drawbacks. For example, both models classify examinees into speeded or nonspeeded groups, and estimate nonspeeded parameters using only a subset of the data. Also, by assuming that speededness only manifests itself in random guessing, the hybrid model is likely unrealistic. The mixture model approach, on the other hand, is sensitive to examinees whose performance on end-of-test items is appreciably worse than on the rest of the test; therefore, it requires examinees to have achieved a certain level of performance prior to becoming speeded. Consequently, the mixture model is biased against identifying low-ability speeded examinees. The mixture model approach is also extremely time-consuming. More importantly, however, is that testing companies in the United States may not be allowed to use the mixture model for purposes of reporting scores to examinees. Under Title I of the Civil Rights Act of 1991 (1991), it is illegal to use different cut-scores for different manifest groups of test takers. Though the mixture model has not been

subjected to litigation, it is unclear whether it would be deemed permissible to score exams using different item parameters (for the same items) for latent groups of examinees.

Therefore, it would be desirable to have a model that accounts for speededness and can overcome some of the limitations with current models. In spite of the success the 3PLt has had in accounting for other types of LID, the model has not previously been studied in the speededness context. In this study, we consider a 3PLt mixed model (3PLt*), where the items early in the test are assumed to be locally independent and are modeled by the 3PLM, but items at the end of the test are assumed to be speeded and are modeled by the 3PLt. The purpose of this study was to examine the utility of the 3PLt*, under simulated conditions of test speededness, via comparison with both the traditional 3PLM and a mixture three-parameter logistic model (M3PLM) for test speededness (Bolt et al., 2002, 2003).

Research Design

Data Simulation

Item responses were generated using a model for speeded test data that allows for speeded examinees' performances to decay at different rates and times (Goegebeur, DeBoeck, Wollack, & Cohen, conditional acceptance; Wollack & Cohen, 2004). This model is given by:

$$P_i^*(\theta_j) = c_i + (1-c_i)P_i(\theta_j)\min\left\{1, \left[1-\left(\frac{i}{n}-\eta_j\right)\right]^{\lambda_j}\right\} \quad ; \qquad (1)$$

where $P_i(2_j)$ is the standard two-parameter logistic model, $P_i(\theta_j) = \frac{1}{1+e^{-a_i(\theta_j - b_i)}}$, such that a_i and b_i are the item discrimination and difficulty, respectively, for item i (i = 1, . . . , n), c_i is the pseudo-guessing parameter for item i, 2_j is the trait level of examinee j (j = 1, . . . , N), 0_j (0 # 0_j # 1) and 8_j (8_j $ 0) are the speededness point and rate parameters for examinee j, and min [x, y] is the smaller of the two values x and y. In the above model, 0_j models the point in the test,

expressed as the percentage of the items that have been completed, at which examinee *j* first experiences speededness. The θ_j parameter controls the rate at which examinee *j*'s performance deteriorates. The model in Equation (1) is appealing because when $\theta_j = 0$ and/or $\omega_j = 1$, examinee responses are modeled by the 3PLM. However, as θ_j increases and/or ω_j decreases (indicating that the examinee is greatly influenced by speededness), the probability of answering correctly begins to shrink, eventually approaching c_i, asymptotically.

Examinee parameters from a 3PLM were generated randomly from various distributions. Examinee ζ_j parameters were sampled randomly from a $\mathcal{N}(0, 1)$ distribution, for all examinees. However, θ_j and ω_j parameters were sampled differently for speeded and nonspeeded examinees. For nonspeeded examinees, θ_j and ω_j were fixed at 0 and 1, respectively, so responses were generated from the 3PLM. For speeded examinees, however, coefficient λ_j^* was sampled randomly from a $\mathcal{N}(3.5, 1)$, and θ_j was computed as $\theta_j = exp(\lambda_j^*)$. ω_j parameters were generated from two different distributions, either a Beta (18, 2) or a Beta (16, 4), simulating tests with different amounts of speededness. In the former case, speeded examinees, on average, became speeded after having completed 90% of the test items (i.e., on average, they were speeded for the last 6 items). In the latter case, the average speeded examinee began to experience speededness effects after 80% of the test was complete (i.e., on the final 12 items).

3PLM item parameters for a 60-item test were fixed, so that $a_i = 1$, $b_i = 0$, and $c_i = 0.2$ for all items. Although item parameters in simulation studies are typically randomly sampled from representative distributions, the current simulation design was selected to ensure that end-of-test items were no more difficult than other items, at least for the nonspeeded group. Having difficult items at the end of the test can mask speededness effects by making it difficult to discern

whether examinees are attempting and missing those questions or whether their performance has deteriorated due to speededness. By constraining the item parameters to be equal for all items, any changes in examinee behavior may be attributed solely to speededness effects.

Item responses for 1,500 nonspeeded examinees and 500 speeded examinees were generated from Equation (1) using the above-specified parameters. For convenience, examinees 1 – 1,500 were simulated to be nonspeeded and examinees 1,501 – 2,000 were simulated as speeded. Five different datasets were generated for both speededness conditions. All person parameters (i.e., θ_j, τ_j, and ω_j) were resampled from their respective distributions for each replication.

Estimation Models

Each dataset was analyzed by three separate models: 3PLt*, 3PLM, and M3PLM. In estimating the parameters of the 3PLt* and M3PLM, it is necessary to specify how end-of-test items will be modeled. For the 3PLt*, the potentially speeded end-of-test items were chunked into one or more testlets, while the remaining items (at the beginning and middle of the test) were assumed to satisfy the LID assumption, hence were modeled by the 3PLM. Under the M3PLM, end-of-test items were modeled to have two distinct sets of b_i values such that $b_{i,1} \geq b_{i,2}$ for all $i \geq n'$, where n' is the first end-of-test item. For all $i < n'$, b_i values were constrained so that $b_{i,1} = b_{i,2}$. Equality constraints were placed on a_i and c_i for all items. Estimation of all models was done in WinBugs (Spiegelhalter, Thomas, & Best, 2000), using a Markov chain Monte Carlo algorithm (MCMC; Gilks, Richardson, & Spiegelhalter, 1996; Patz & Junker, 1999a, 1999b). Appropriate burn-ins for the different models were determined from pilot runs. In the case of the 3PLt* and 3PL models, the initial 1,000 iterations were discarded. For the M3PLMs, the initial 4,000 iterations were discarded. For all models, a minimum of 5,000 iterations were sampled after burn-in. The average sampled value across all iterations after burn-

in was taken as the parameter estimate.

Given that the only source of LID simulated was test speededness, it was expected that γ_{jk} values would be negligible for nonspeeded examinees but substantial for speeded examinees. However, under the 3PLt*, γ_{jk} parameters are estimated subject to the constraint $\sum_j \hat{\gamma}_{jk} = 0$, for each of the *k* testlets. As a result of centering the $\hat{\gamma}_{jk}$ across both speeded and nonspeeded examinees, the $\hat{\gamma}_{jk}$ values for nonspeeded examinees were largely negative, whereas they tended to be positive for the speeded examinees. Therefore, the item difficulty estimates were shifted away from where they would have been had the model parameters been estimated using only nonspeeded examinees. To compensate for this bias, item difficulties for end-of-test items were adjusted by a constant that reflects the mean $\hat{\gamma}_{jk}$ for nonspeeded examinees. This process is described below.

If *q* and *q*′ denote the nonspeeded and speeded examinees, respectively, the appropriate correction factor to be added to each of the b_i is equal to $\overline{\gamma}_{qk}$, the average γ_{jk} across all nonspeeded examinees. However, in practice, neither the γ_{jk} nor the exact subset comprising *q* is known; therefore the value $\overline{\gamma}_{qk}$ must be estimated. Because speededness is assumed to be the only source of LID, it is expected that $\gamma_{q,k} < \gamma_{q',k}$ for all (*q*, *q*′) pairs. Therefore, under the assumption that the $\gamma_{jk} \sim \mathcal{N}(0, \sigma^2_{\gamma_k})$, an approximation for $\overline{\gamma}_{qk}$ is given by $\hat{\overline{\gamma}}_{qk} = \hat{\sigma}_{\gamma_k} \mathcal{E}\left(Z^T_{1-\frac{q}{q+q'}} \right)$, where $q / (q + q')$ is the proportion of examinees who are nonspeeded, $\hat{\sigma}^2_{\gamma_k}$ is the teslet variance estimated within the MCMC algorithm, and $\mathcal{E}\left(Z^T_{1-\frac{q}{q+q'}} \right)$ is the mean of the standard normal distribution truncated on the left at $\alpha = 1 - q / (q + q)$. Here, because it was known that three-quarters of the sample was nonspeeded examinees, $\mathcal{E}\left(Z^T_{1-\frac{q}{q+q'}} \right) = \mathcal{E}\left(Z^T_{0.25} \right) = -0.4242$ was used for all testlets and datasets.

In addition, in order to estimate the (parameters in the 3PLt*, it is necessary to specify the size of the testlets. In situations where LID is caused by dependency on a common stimulus,

specification of the testlet size is easy, as all items associated with that stimulus are analyzed as a testlet. In the context of speededness, specification of testlet size is more difficult for two reasons. First, the stimulus common to all affected items does not have a clear, discernable beginning. Some examinees will be speeded, others will not be. Further, the speeded examinees do not all become speeded at the same point. Therefore, it is unclear how many items should comprise the testlet(s) at the end of the test. The second problem is that, even if it were knowable which items were affected by speededness, the effect of LID may not be constant across all speeded items. Instead, speededness may become more pronounced, resulting in a higher dependence among the items at the very end of the test than other speeded items somewhat earlier in the test. Yet, within a testlet, the 3PLt γ parameter exerts a constant effect on item difficulty estimates.

The first problem above—how to identify the end-of-test items—also exists for the M3PLM, in that the model requires specification of the items that are constrained to be harder for the speeded class. Traditionally, the constraints have been placed on the last 6-10 items, items that are believed to contain the most speededness. Conventional wisdom is that sorting examinees into classes accurately requires only that the most heavily speeded items be considered. However, if the set of end-of-test items is too small, classifications based on those items may be unreliable. If the set is too large, ordinal constraints will be imposed on items which do not behave differently for the two groups, resulting in difficulty sampling acceptable values for the MCMC estimation. The issue of how to select end-of-test items for speededness analyses has not been studied empirically.

To explore these issues further, parameters for the 3PLt* and M3PLM were estimated according to different specifications. For each dataset, 3PLt* parameters were estimated six

times, manipulating both the number of items assumed to be speeded (i.e., end-of-test items) and the number of testlets. The estimation specifications for the end-of-test items with the 3PLt* are shown in Table 1. All remaining (non end-of-test) items were estimated with the 3PLM. For the M3PLM, each dataset was fitted with three different mixture models, varying the number of end-of-test items assumed to be speeded. The M3PLM was estimated by constraining the final 4, 8, and 16 items to be harder for the speeded class. Under these three models, b_i values for the initial 56, 52, and 44 items, respectively, were constrained to be equal for the latent speeded and nonspeeded groups.

Insert Table 1 About Here

Evaluative Measures

To assess the quality of the item and ability parameter recovery, root mean square errors (RMSE) and biases were computed between generating parameters and their estimates for all models. Prior to computing RMSEs and biases, item parameter estimates for all replications were equated to the metric of the generating item parameters using the test characteristic curve method (Stocking & Lord, 1983), as implemented in the EQUATE computer program (Baker, Al-Karni, & Al-Dosary, 1991). All 60 items were included in the anchor set. For the M3PLM, the equating transformation coefficients were estimated from the item parameter estimates from the nonspeeded class only. It is important to note that, for the 3PLt*, the quadratic loss function to be minimized in characteristic curve equating would typically also include estimates of the means of the estimated testlet factors (μ_γ , Li, Bolt, & Fu, 2005). However, because the data were not simulated as 3PLt data, true μ_γ values do not exist. Therefore, the μ_γ were not included in estimating the equating coefficients for the 3PLt*.

Parameter recovery was further assessed by computing correlations between estimated and generating ability values. Because item parameters were identical for all items, there was zero variance among the generating values, so correlations could not be computed.

In addition, a number of statistics were computed to determine the overall fit of the model. After fitting the model, Yen's Q_3 (1984) was computed to assess the amount of LID between pairs of items for which the model does not account. Q_3 is the average correlation between item residuals (i.e., observed item score minus expected item score) for pairs of items. Because speededness causes LID at the end of the test, Q_3 statistics were computed for the entire test, as well as separately for the last 4, 8, 12, and 16 items. Further, to explore how well the 3PLt* accounted for end-of-test LID caused by test speededness, we looked at two variables related to the γ_{jk} for examinee j on testlet k. Estimated testlet variances ($\hat{\sigma}^2_{\gamma_k}$) were computed directly from the WinBugs runs for each of the 3PLt* models. Because under the generating model, the amount of deterioration in performance increases as a speeded examinee progresses through the test, LID should increase throughout the items at the end of the test. Consequently, it was expected that $\hat{\sigma}^2_{\gamma_k}$ would increase from the first to the last testlet. In addition, because LID was not simulated for nonspeeded examinees, it was expected that (a) within testlet k, the average $\gamma_{\bullet k}$ value would be greater among speeded examinees than nonspeeded examinees, and (b) across testlets, the difference in average testlet values between speeded and nonspeeded groups (i.e., $\mu_{\gamma_{k,Speeded}} - \mu_{\gamma_{k,Nonspeeded}}$) would increase from the first to the last testlet. Therefore, $\hat{\sigma}^2_{\gamma_k}$, $\overline{\gamma}_{k,Speeded}$, and $\overline{\gamma}_{k,Nonspeeded}$ values were examined to learn the extent to which their values conformed to expectations.

Results

Biases and RMSEs for ability parameters under all the models and both magnitudes of speededness are provided in Table 2, averaged across replications. Separate biases and RMSEs are provided for the nonspeeded (NS) and speeded (SP) examinees and the total sample. Average biases and RMSEs for item parameters are provided in Table 3 (for b_i values) and Table 4 (for a_i values). RMSEs and biases were uniformly very low for pseudo-guessing parameters, so results are not presented here. Statistics are presented separately for NS and SP items. Because the expected number of speeded items were 6 and 12 in the $\mathcal{E}(\eta) = 0.90$ and $\mathcal{E}(\eta) = 0.80$ conditions, respectively, items 55-60 were considered SP and in the $\mathcal{E}(\eta) = 0.90$ condition and items 49-60 were considered SP and in the $\mathcal{E}(\eta) = 0.80$ condition, regardless of the number of items actually analyzed as end-of-test items under the different models.

Insert Table 2 About Here

Regardless of the magnitude of speededness, all models showed very small overall (i.e., total) biases in ability estimation. The largest absolute bias, just 0.04, was for the 3PLt*-2×4 and 3PLt*-1×4 models in the $\mathcal{E}(\eta) = 0.80$ condition. Most of the models had overall biases that were no more than 0.02 in absolute value. The NS ability biases for all models were positive (i.e., ability was over-estimated), regardless of the amount of speededness, and were quite similar, with the possible exception of the 3PLt*-4×4 model in the $\mathcal{E}(\eta) = 0.80$ condition which produced a slightly larger bias. In both conditions, for the SP group, biases were consistently negative for all models (i.e., ability was under-estimated). In the $\mathcal{E}(\eta) = 0.90$ condition, SP bias was largest for the 3PLM; SP biases for the M3PLMs and 3PLt* models were similar. In the $\mathcal{E}(\eta) = 0.80$

condition, however, SP biases were lowest for the M3PLMs and were worst for the 3PLM. Performance of the 3PLt* models varied depending on the number of end-of-test items that were modeled with testlets. The 3PLt*-1×16, 3PLt*-2×8, and 3PLt*-4×4, each of which used testlets to describe the performance on the final 16 items, performed similarly and consistently produced SP biases most similar to that in the M3PLMs. The two 8-item 3PLt* models (i.e., 3PLt*-1×8, and 3PLt*-2×4) produced the next lowest biases, followed by the 3PLt*-1×4, which was very similar to the 3PLM. For both the16- and 8-item 3PLt* models, bundling end-of-test items into fewer testlets resulted in lower SP biases than using more testlets.

RMSEs for NS, SP, and overall were similar for all models in the $\mathcal{E}(\eta) = 0.90$ condition. When $\mathcal{E}(\eta) = 0.80$, the RMSE pattern seemed to coincide with the bias pattern. RMSEs for the M3PLMs were slightly lower than for the 3PLt* models, and among the 3PLt* models, RMSEs were lowest when 16 items were modeled with testlets. The 3PLM had the highest RMSE, but the 3PLt* that modeled only 4 testlet items was not much better.

Overall, the M3PLM seemed to recover ability parameters best. This was most noticeable among the SP bias in the $\mathcal{E}(\eta) = 0.80$ condition. Results from the three M3PLM solutions were virtually indistinguishable from one another. Results from the six 3PLt* solutions varied, but these models tended to outperform the 3PLM, particularly with respect to SP bias. Among the 3PLt* models, recovery of ability parameters appeared to suffer most when the testlets were not sufficiently large to include most of the speeded items. This makes sense because items that are not included in testlets are modeled by the 3PLM, where performance was clearly worst.

Bias and RMSE results for item difficulty estimates (see Table 3) were largely similar to those for ability estimates. The biases and RMSEs for the M3PLMs were small for all groups of items in both conditions, indicating that the M3PLMs recovered underlying difficulty parameters well for both speeded and nonspeeded items. In the $\mathcal{E}(\eta) = 0.90$ condition, the 3PLt* models worked very well at recovering item difficulty estimates for both speeded and nonspeeded items,

except for in the 3PLt*-1×16 model, where bias was fairly large. In the $\mathcal{E}(\eta) = 0.80$ condition, the 3PLt* difficulty estimates for speeded items were somewhat over-estimated, relative to their estimates in the M3PLMs. This was especially true in the 3PLt*-1×4.

Both the M3PLMs and 3PLt* models worked significantly better than the 3PLM at recovering item difficulties. Under the 3PLM, item difficulties for end-of-test items were, on average, estimated as .32 logits too hard in the $\mathcal{E}(\eta) = 0.90$ condition and 0.35 logits too hard in the $\mathcal{E}(\eta) = 0.80$ condition. In addition, RMSEs for the SP items were much higher than for the other models.

Insert Table 3 About Here

The data in Table 4 suggests that both the M3PLM and 3PLt* were also better at recovering the underlying item discrimination indices than was the 3PLM. The difference was not as apparent in the $\mathcal{E}(\eta) = 0.90$ condition, where biases for all models were fairly comparable overall. In fact, the 3PLt* and 3PLM actually produced slightly less biased and less variable discrimination estimates among end-of-test items than did the M3PLM. However, in the $\mathcal{E}(\eta) = 0.80$ condition, the 3PLM over-estimated end-of-test discrimination parameters by an average of 0.27. In contrast, bias among the other models ranged from just -0.01 to 0.08. RMSEs for the 3PLM were also two-and-a-half to three times larger than for either the M3PLM or 3PLt* models.

Insert Table 4 About Here

Correlations between generating and estimated θ parameters are provided in Table 5 for the different models. In the $\mathcal{E}(\eta) = 0.90$ condition, correlations for the three groups were high for all

models. The correlations were slightly lower when $\mathcal{E}(\eta) = 0.80$, particularly among the speeded examinees. However all correlations studied were at least .90 and within both conditions, differences between the models were negligible.

Insert Table 5 About Here

In addition to examining the quality of parameter recovery, analyses were performed to determine the extent to which the models accounted for LID. Table 6 shows the average Q_3 statistics for all models under both speededness conditions for the entire test, as well as for the last 4, 8, 12, and 16 items. Although Q_3 should be distributed approximately $\mathcal{N}(0, 1/(N-3))$, Yen (1993) showed that, in practice, Q_3 has a slight negative bias equal to $-1/(n-1)$. Therefore, average Q_3 values near (or just below) zero are indicative of item pairs that are free from LID.

Insert Table 6 About Here

Over all item pairs, Q_3 values were low for all models in both speededness conditions. However, in large part, this is because the nonspeeded items were locally independent, and the number of nonspeeded items outweighed the number of speeded items by an average of at least 4:1, even in the high speededness condition. Furthermore, the number of nonspeeded examinees was three times greater than the number of speeded examinees.

The ability of the models to remove LID due to speededness is best assessed by considering the average Q_3 indices among pairs of items at the end of the test. From Table 6, one can see that the models were not all equally effective at accounting for speededness. The three M3PLMs were very effective at removing LID from the datasets. Regardless of $\mathcal{E}(\eta)$, average Q_3 indices

were essentially zero between pairs of end-of-test items. Although the 3PLt* was not as effective as the M3PLM, Table 6 shows that it did effectively reduce the amount of LID that would have been present using only the 3PLM. In the $\mathcal{E}(\eta) = 0.90$ condition, after fitting a 3PLt*, there was very little evidence of remaining LID, with the exception of the 3PLt*-1×16 condition, where the average Q_3 value among the final four items was 0.09. It is quite possible that this effect was observed because the number of speeded items (six, on average) was substantially less than the number of items in the last testlet (16). That is, the magnitude of dependency among the last few speeded items was largely offset (i.e., watered down) by including in the testlet many items for which local independence held. Consequently, the $\hat{\gamma}_{jk}$ under-estimated γ_{jk}, so the b_i for end-of-test items were not corrected enough to remove the existing LID.

The $\mathcal{E}(\eta) = 0.80$ condition shows one of the possible limitations of using the 3PLt* in the speededness context. Provided the testlets are appropriately specified, the 3PLt* appears to work rather well at removing LID. Note that the 3PLt*-1×16, 3PLt*-2×8, and 3PLt*-1×8 all do a fairly good job accounting for LID. However, the other three 3PLt* models do not work as well, particularly in terms of LID among the last 8 or last 12 items. Consistent with some patterns that were found earlier, it appears as though fewer, larger testlets is more effective than having more, smaller testlets. Still, from Table 6 it is clear that the 3PLt* models, even at their worst, offer a substantial improvement upon the 3PLM with regards to the amount of LID among end-of-test items.

Finally, testlet variances in the 3PLt* models were examined to see if (a) the variances increased for testlets associated with end-of-test items, and (b) there was greater variation among

$\hat{\gamma}_{jk}$ values for speeded examinees than for nonspeeded examinees. Testlet variances for the various 3PLt* models are shown in Table 7. As can be seen from Table 7, the 3PLt* models were sensitive to the type and magnitude of simulated LID, in that $\hat{\sigma}^2_{\gamma}$ values were small for testlets where little speededness was simulated (e.g., in items 45-52 in the $\mathcal{E}(\eta)$ = 0.90 condition and in items 45-48 in the $\mathcal{E}(\eta)$ = 0.80 condition), and moderate-to-large over portions of the test that were expected to be more heavily speeded. Moreover, the testlet variances tended to be larger when $\mathcal{E}(\eta)$ = 0.80, indicating that the 3PLt* model was able to account for the stronger speededness effect that was simulated.

Insert Table 7 About Here

Table 8 shows the differences between the sample variances of $\hat{\gamma}_{jk}$ ($s^2_{\hat{\gamma}_{jk}}$) for nonspeeded and speeded examinees under the two conditions. Differences were computed as $s^2_{\hat{\gamma}_{k,NS}} - s^2_{\hat{\gamma}_{k,SP}}$, so negative values indicate greater variability in $\hat{\gamma}_{jk}$ among the speeded examinees. Testlets associated with end-of-test items were expected to show LID for only the speeded group. Therefore, on these testlets, $s^2_{\hat{\gamma}_{j,SP}}$ should be higher than $s^2_{\hat{\gamma}_{j,NS}}$, resulting in negative values. In fact, this was precisely the observed trend. For sets of nonspeeded items (e.g., in the 3PLt*-2×8 [testlet 1] and the 3PLt*-4×4 [testlets 1 and 2] when $\mathcal{E}(\eta)$ = 0.90, and in the 3PLt*-4×4 [testlet 1] when$\mathcal{E}(\eta)$ = 0.80), no differences in $s^2_{\hat{\gamma}_{jk}}$ values were observed. However, in all other conditions, $s^2_{\hat{\gamma}_{j,SP}}$ values were larger than $s^2_{\hat{\gamma}_{j,NS}}$ values. The magnitude of difference increased as the test progressed and the extent of speededness increased.

Insert Table 8 About Here

Discussion

This study examined the effectiveness of a 3PLt mixed model at recovering underlying item and ability parameters under conditions of test speededness, through comparison with one model that explicitly models test speededness, the M3PLM, and one model that assumes that no LID exists among the items, the 3PLM. The results of the study showed that while the M3PLM was the most robust to speededness contaimination, the magnitudes of differences in bias and RMSE between the M3PLM and 3PLt* models was typically small, even under heavily speeded conditions. In contrast, the 3PLM produced heavily biased estimates of ability, item difficulty and item discrimination for heavily speeded examinees and failed to account for LID among end-of-test items. Hence, it would appear as though the 3PLt* presents a viable alternative model when a dataset is believed to contain speededness. This is particularly true where trait estimation is concerned, because the legality of estimating ability for a subset of examinees using a different (less stringent) set of item difficulty estimates, as is done with the M3PLM, is questionable.

Not surprisingly, the 3PLt* seemed to work best when the end-of-test items being modeled in testlets coincided reasonably well with the test items that were speeded. When too many items were modeled in testlets, as was the case when $\mathcal{E}(\eta) = 0.90$ and the last 16 items were in testlets, and when too few items were modeled in testlets, as was the case when $\mathcal{E}(\eta) = 0.80$ and the last 4 items were in testlets, parameter recovery was adversely affected. This is particularly apparent by examining the biases in item difficulty and the Q_3 values for end-of-test items, but is present to a certain extent by examining bias in ability, as well. It appears to be the case that, if

the number of speeded test items is roughly known, estimation is improved by including those items into a single testlet. However, if it is unclear whether and how much speededness exists in the dataset, rather than use a single testlet which may over- or under-estimate the number of speeded items, it is preferable to use more, smaller testlets. Concern over the number of items to model as speeded does not appear warranted for the M3PLM: the bias, RMSEs, and Q_3 statistics were all unaffected by the number of end-of-test items specified in the mixture model.

As is the case with all simulation studies, the results of this study, at least to a certain extent, reflect the manner in which the data were simulated. Whereas the M3PLM assumes that examinees belong to one of two discrete classes—either speeded or nonspeeded—the 3PLt* assumes more of a continuum in how it accommodates speededness since the testlet random effect parameter is continuous. In this study, although examinees were either simulated as speeded (i.e., $\delta_j = 0$ and/or $\theta_j = 1$) or nonspeeded (i.e., $\delta_j > 0$ and $0 \leq \theta_j < 1$), those who were speeded were simulated to vary in both degree and amount. Therefore, the simulation of speededness was perhaps more consistent with the way it is modeled under the 3PLt* than with either of the other two models. To the extent that this simulation did not reflect actual patterns of speededness in real data, it may not be possible to generalize these findings.

The testlet model has historically been used in situations where the nature of the tasks creates clearly defined sets of items that are somehow related, as is often the case with reading comprehension questions associated with a common passage. The results of this study open the door to explore the use of the testlet model in other, more complicated settings for which items contain LID. For example, it would be worthwhile to explore whether the testlet model might be applicable for situations where multiple types of LID are simultaneously present, such as with a partially speeded test of reading comprehension. In addition, future study should explore using

the testlet model to compensate for possible LID that exists among less clearly defined sets of items, such as sets of items that might be speeded. Other examples of less clearly defined sets worth studying include items grouped by type, format, or objective. In addition, it would be worthwhile to investigate whether the testlet approach could be extended to accommodate a progressive decline in performance as a result of test speededness, such as was simulated in this study. These extensions and further explorations are left to future work.

References

Baker, F. B., Al-Karni, A., & Al-Dosary, I. M. (1991). EQUATE: A computer program for the test characteristic curve method of IRT equating. *Applied Psychological Measurement, 50,* 529-549.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Applications of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement, 39,* 331-348.

Bolt, D. M., Mroch, A. A., & Kim, J.-S. (April, 2003). *An empirical investigation of the Hybrid IRT model for improving item parameter estimation in speeded tests.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

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

Civil Rights Act of 1991, Pub. L. No. 102-166, §106 (1991).

Douglas, J., Kim, H. R., Habing, B., & Gao, F. (1998). Investigating local dependence with conditional covariance functions. *Journal of Educational and Behavioral Statistics, 23*, 129-151.

Du, Z. (1998). Modeling conditional item dependencies with a three-parameter logistic testlet model. *Dissertation Abstracts International, 59*(10), 5429. (UMI No. 9910577)

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J., Eds. (1996). *Markov chain Monte Carlo in practice*. London: Chapman & Hall.

Goegebeur, Y., DeBoeck, P., Wollack, J. A., & Cohen, A. S. (conditional acceptance). A speeded item response model with gradual process change. *Psychometrika*.

Lee, G., Kolen, M. J., Frisbie, D. A., & Ankenmann, R. D. (2001). Comparison of dichotomous and polytomous item response models in equating scores from tests composed of testlets. *Applied Psychological Measurement, 25*, 357-372.

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

Li, Y. & Cohen, A. S. (April, 2003). *Equating tests composed of testlets: A comparison of a testlet response model and four polytomous response models*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement, 31*, 200-219.

Patz, R. J., & Junker, B. W. (1999a). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24*, 146-178.

Patz, R. J., & Junker, B. W. (1999b). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics, 24*, 342-366.

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

Spiegelhalter, D. J., Thomas, A., & Best, N. G. (2000). *WinBUGS version 1.3* [Computer program]. Robinson Way, Cambridge CB2 2SR, UK: Institute of Public Health, Medical Research Council Biostatistics Unit.

Stocking, M., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 207-210.

Thissen, D., Steinberg, L., & Mooney, J. (1989). Trace lines for testlets: A use of multiple-categorical response models. *Journal of Educational Measurement, 26*, 247-260.

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.

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: Issues and Practice, 15,* 22-29.

Wollack, J. A., & Cohen, A. S. (2004, April). *A model for simulating speeded test data*. Presentation at the annual meeting of the American Educational Research Association, San Diego, CA.

Wollack, J. A., Cohen, A. S., & Wells, C. S. (2003). A method for maintaining scale stability in the presence of test speededness. *Journal of Educational Measurement, 40,* 307-330.

Yamamoto, K. & Everson, H. (1997). Modeling the effects of test length and test time on parameter estimation using the HYBRID model. In J. Rost & R. Langeheine (Eds.) *Applications of Latent Trait and Latent Class Models in the Social Sciences.* New York: Waxmann.

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.

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

Table 1

Explanation of the number and size of testlets for the 3PLt

Model Abbreviation	Number of Teslets at End of Test	Number of Items Per Testlet	Total Number of Items Treated as Speeded
3PLt*-1×16	1	16	16
3PLt*-2×8	2	8	16
3PLt*-4×4	4	4	16
3PLt*-1×8	1	8	8
3PLt*-2×4	2	4	8
3PLt*-1×4	1	4	4

Table 2. Bias and RMSE for Ability Parameters

	Bias			RMSE		
	NS	SP	Total	NS	SP	Total
$\mathcal{E}(\eta) = 0.90$						
M3PLM-4	0.03	-0.03	0.02	0.51	0.53	0.51
M3PLM-8	0.03	-0.04	0.01	0.51	0.54	0.52
M3PLM-16	0.03	-0.04	0.01	0.51	0.53	0.52
3PLt-1×16	0.03	-0.05	0.01	0.53	0.56	0.53
3PLt-2×8	0.04	-0.03	0.02	0.52	0.55	0.53
3PLt-4×4	0.04	-0.04	0.02	0.52	0.55	0.53
3PLt-1×8	0.03	-0.04	0.02	0.52	0.55	0.53
3PLt-2×4	0.03	-0.05	0.01	0.52	0.55	0.53
3PLt-1×4	0.02	-0.06	0.00	0.52	0.55	0.53
3PLM	0.01	-0.09	-0.01	0.51	0.56	0.52
$\mathcal{E}(\eta) = 0.80$						
M3PLM-4	0.03	-0.03	0.02	0.51	0.54	0.52
M3PLM-8	0.03	-0.03	0.02	0.51	0.54	0.52
M3PLM-16	0.04	-0.02	0.02	0.51	0.55	0.52
3PLt-1×16	0.04	-0.05	0.02	0.52	0.58	0.54
3PLt-2×8	0.04	-0.08	0.01	0.51	0.58	0.53
3PLt-4×4	0.06	-0.10	0.02	0.50	0.59	0.53
3PLt-1×8	0.03	-0.11	-0.01	0.51	0.59	0.53
3PLt-2×4	0.00	-0.16	-0.04	0.50	0.61	0.53
3PLt-1×4	0.01	-0.18	-0.04	0.50	0.63	0.54
3PLM	0.02	-0.19	-0.03	0.50	0.65	0.54

Note: NS refers to simulated nonspeeded examinees, SP refers to simulated speeded examinees.

Table 3. Bias and RMSE for Item Difficulty Parameters

	Bias			RMSE		
	NS	SP	Total	NS	SP	Total
$\mathcal{E}(\eta) = 0.90$						
M3PLM-4	0.04	0.04	0.04	0.08	0.09	0.08
M3PLM-8	0.04	0.04	0.04	0.08	0.09	0.08
M3PLM-16	0.04	0.02	0.04	0.08	0.09	0.09
3PLt-1×16	0.03	0.14	0.04	0.06	0.15	0.08
3PLt-2×8	0.05	0.07	0.05	0.07	0.10	0.08
3PLt-4×4	0.05	0.02	0.05	0.07	0.07	0.07
3PLt-1×8	0.05	0.06	0.05	0.07	0.10	0.08
3PLt-2×4	0.05	0.01	0.04	0.06	0.06	0.06
3PLt-1×4	0.04	0.05	0.04	0.06	0.10	0.06
3PLM	0.02	0.32	0.05	0.07	0.34	0.12
$\mathcal{E}(\eta) = 0.80$						
M3PLM-4	0.06	0.01	0.05	0.11	0.09	0.11
M3PLM-8	0.05	0.01	0.05	0.09	0.08	0.09
M3PLM-16	0.05	0.01	0.05	0.09	0.07	0.09
3PLt-1×16	0.05	0.08	0.06	0.08	0.10	0.09
3PLt-2×8	0.05	0.04	0.05	0.07	0.08	0.07
3PLt-4×4	0.06	0.06	0.06	0.08	0.09	0.08
3PLt-1×8	0.04	0.06	0.04	0.06	0.13	0.08
3PLt-2×4	0.01	0.07	0.02	0.05	0.13	0.07
3PLt-1×4	0.01	0.16	0.04	0.05	0.20	0.10
3PLM	-0.01	0.35	0.06	0.08	0.36	0.17

Note: NS refers to nonspeeded items, SP refers to speeded items. In the $\mathcal{E}(\eta) = 0.90$ condition, the last 6 items (items 55-60) were considered SP and in the $\mathcal{E}(\eta) = 0.80$ condition, the last 12 items (items 49-60) were considered SP, regardless of the number of items actually analyzed as end-of-test items under the different models

Table 4. Bias and RMSE for Item Discrimination Parameters

	Bias			RMSE		
	NS	SP	Total	NS	SP	Total
$\mathcal{E}(\eta) = 0.90$						
M3PLM-4	0.03	0.12	0.04	0.10	0.19	0.11
M3PLM-8	0.03	0.12	0.04	0.10	0.19	0.12
M3PLM-16	0.03	0.11	0.04	0.11	0.20	0.12
3PLt-1×16	0.05	-0.08	0.04	0.09	0.12	0.09
3PLt-2×8	0.05	-0.01	0.04	0.09	0.09	0.09
3PLt-4×4	0.05	-0.03	0.04	0.09	0.10	0.09
3PLt-1×8	0.05	-0.01	0.05	0.10	0.09	0.09
3PLt-2×4	0.05	-0.04	0.04	0.09	0.11	0.09
3PLt-1×4	0.05	-0.04	0.04	0.09	0.10	0.09
3PLM	0.05	-0.04	0.04	0.11	0.12	0.11
$\mathcal{E}(\eta) = 0.80$						
M3PLM-4	0.04	0.06	0.05	0.12	0.13	0.12
M3PLM-8	0.04	0.06	0.04	0.12	0.13	0.12
M3PLM-16	0.04	0.06	0.04	0.12	0.13	0.12
3PLt-1×16	0.06	0.04	0.05	0.12	0.14	0.12
3PLt-2×8	0.05	-0.01	0.04	0.09	0.11	0.09
3PLt-4×4	0.04	0.08	0.05	0.09	0.14	0.10
3PLt-1×8	0.05	-0.01	0.04	0.09	0.10	0.09
3PLt-2×4	0.04	0.08	0.05	0.09	0.14	0.10
3PLt-1×4	0.04	0.07	0.05	0.09	0.15	0.10
3PLM	0.02	0.27	0.07	0.11	0.37	0.19

Note: NS refers to nonspeeded items, SP refers to speeded items. In the $\mathcal{E}(\eta) = 0.90$ condition, the last 6 items (items 55-60) were considered SP and in the $\mathcal{E}(\eta) = 0.80$ condition, the last 12 items (items 49-60) were considered SP, regardless of the number of items actually analyzed as end-of-test items under the different models

Table 5. Average Correlations Between Estimated and Generating Ability Parameters

	Average Correlations		
	NS	SP	Total
$\mathcal{E}(\eta) = 0.90$			
M3PLM-4	0.94	0.93	0.93
M3PLM-8	0.93	0.91	0.92
M3PLM-16	0.93	0.92	0.93
3PLt-1×16	0.94	0.93	0.93
3PLt-2×8	0.94	0.93	0.93
3PLt-4×4	0.94	0.93	0.93
3PLt-1×8	0.94	0.93	0.93
3PLt-2×4	0.94	0.93	0.93
3PLt-1×4	0.94	0.93	0.93
3PLM	0.94	0.93	0.93
$\mathcal{E}(\eta) = 0.80$			
M3PLM-4	0.93	0.91	0.92
M3PLM-8	0.93	0.91	0.92
M3PLM-16	0.93	0.91	0.93
3PLt-1×16	0.93	0.91	0.92
3PLt-2×8	0.93	0.91	0.92
3PLt-4×4	0.94	0.91	0.92
3PLt-1×8	0.93	0.91	0.92
3PLt-2×4	0.94	0.91	0.92
3PLt-1×4	0.94	0.91	0.92
3PLM	0.94	0.90	0.91

Note: NS refers to simulated nonspeeded examinees, SP refers to simulated speeded examinees.

Table 6. Average Q_3 Statistics for All Items and End-of-Test Items

	Average Q_3				
	Average	Last 16	Last 12	Last 8	Last 4
$\mathcal{E}(\eta) = 0.90$					
M3PLM-4	0.01	0.00	0.00	0.00	0.00
M3PLM-8	0.01	0.00	0.00	0.00	0.00
M3PLM-16	0.01	0.01	0.01	0.01	0.01
3PLt-1×16	0.01	0.00	0.00	0.03	0.09
3PLt-2×8	0.01	0.00	0.00	-0.02	0.03
3PLt-4×4	0.01	0.00	0.00	0.01	-0.04
3PLt-1×8	0.01	0.00	0.00	-0.02	0.03
3PLt-2×4	0.01	0.01	0.00	0.01	-0.04
3PLt-1×4	0.01	0.01	0.01	0.02	-0.04
3PLM	0.01	0.02	0.03	0.06	0.09
$\mathcal{E}(\eta) = 0.80$					
M3PLM-4	0.01	0.00	0.00	0.00	0.00
M3PLM-8	0.01	0.00	0.00	0.00	0.00
M3PLM-16	0.01	0.00	0.00	0.00	0.00
3PLt-1×16	0.00	0.00	0.01	0.03	0.04
3PLt-2×8	0.00	0.01	0.02	-0.02	-0.01
3PLt-4×4	0.00	0.02	0.04	0.05	-0.04
3PLt-1×8	0.00	0.02	0.03	-0.02	-0.01
3PLt-2×4	0.00	0.03	0.05	0.05	-0.03
3PLt-1×4	0.00	0.03	0.06	0.08	-0.03
3PLM	0.00	0.07	0.09	0.10	0.10

Table 7. Estimated Testlet Variances

	$\hat{\sigma}^2_{\gamma_1}$	$\hat{\sigma}^2_{\gamma_2}$	$\hat{\sigma}^2_{\gamma_3}$	$\hat{\sigma}^2_{\gamma_4}$
$\mathcal{E}(\eta) = 0.90$				
3PLt-1×16	0.17			
3PLt-2×8	0.09	0.65		
3PLt-4×4	0.12	0.14	0.24	1.37
3PLt-1×8	0.65			
3PLt-2×4	0.29	1.30		
3PLt-1×4	1.26			
$\mathcal{E}(\eta) = 0.80$				
3PLt-1×16	0.74			
3PLt-2×8	0.24	1.24		
3PLt-4×4	0.19	0.49	1.07	1.15
3PLt-1×8	1.20			
3PLt-2×4	1.03	1.10		
3PLt-1×4	1.00			

Table 8. Differences Between Estimated Testlet Variances for Nonspeeded and Speeded Examinees

	$\hat{\sigma}^2_{\gamma_{1,NS}} - \hat{\sigma}^2_{\gamma_{1,SP}}$	$\hat{\sigma}^2_{\gamma_{2,NS}} - \hat{\sigma}^2_{\gamma_{2,SP}}$	$\hat{\sigma}^2_{\gamma_{3,NS}} - \hat{\sigma}^2_{\gamma_{3,SP}}$	$\hat{\sigma}^2_{\gamma_{4,NS}} - \hat{\sigma}^2_{\gamma_{4,SP}}$
$\mathcal{E}(\eta) = 0.90$				
3PLt-1×16	-0.14			
3PLt-2×8	0.01	-0.53		
3PLt-4×4	0.01	0.01	-0.08	-0.77
3PLt-1×8	-0.53			
3PLt-2×4	-0.09	-0.77		
3PLt-1×4	-0.76			
$\mathcal{E}(\eta) = 0.80$				
3PLt-1×16	-0.86			
3PLt-2×8	-0.13	-1.13		
3PLt-4×4	-0.01	-0.22	-0.66	-0.75
3PLt-1×8	-1.10			
3PLt-2×4	-0.63	-0.72		
3PLt-1×4	-0.67			