Using the Testlet Model to Mitigate Test Speededness Effects

James A. Wollack
Youngsuk Suh

University of Wisconsin – Madison

April 9, 2006

RUNNING HEAD: Mitigating Speededness Effects

Using the Testlet Model to Mitigate Test Speededness Effects

This paper studies the effectiveness of the three-parameter testlet 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 the M3PLM worked very well at recovering the underlying parameters, whereas both the 3PLt and 3PLM estimates were biased, particularly for end-of-test items and speeded examinees. Although the 3PLt recovered item parameters better than the 3PLM, there was no appreciable difference between the models for ability estimation.

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. The purpose of this study is to compare the item and ability parameter recovery of the 3PLt with the traditional 3PLM and a mixture three-parameter logistic model (M3PLM) for test speededness (Bolt et al., 2002, 2003), when data are simulated to include varying amounts of test speededness. The effects of different numbers and sizes of testlets were studied, as were the number of modeled speeded items in the M3PLM.

Research Design

Data Simulation

Item responses were generated using a model for generating realistic speeded test data (Goegebeur, DeBoeck, Wollack, & Cohen, conditional acceptance; Wollack & Cohen, 2004). This model is given by:

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

where $P_i(\theta_j)$ 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, θ_j is the trait level of examinee j (j = 1, . . . , N), η_j (0 # η_j # 1) and λ_j (λ_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, η_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, asymptotically, as θ_j increases and/or ω_j decreases (indicating that the examinee is greatly influenced by speededness), the probability of answering correctly approaches c_i. As θ_j decreases towards 0 and/or ω_j increases towards 1, however, this model approaches the 3PLM.

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. θ_j and ω_j parameters, however, 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., were speeded for the last 6 items). In the latter case, the average speeded examinee became speeded 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. Five different datasets were generated for both speededness conditions. All person parameters (i.e., 2_j, 8_j, and 0_j) were resampled from their respective distributions for each replication.

Estimation Models

Each dataset was analyzed by three separate models: 3PLM, 3PLt, 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 end-of-test items were chunked into one or more testlets, while the remaining items (at the beginning and middle of the test) constituted a single, separate testlet. 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 (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.

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.

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 treated as a single testlet. 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. (Throughout the rest of this paper, these three models will be abbreviated as M3PLM-4, M3PLM-8, and M3PLM-16, respectively.) 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 recovery, root mean square errors (RMSE) and biases were computed between generating item and ability 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 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 parameter estimates from the nonspeeded class only. Those coefficients were then applied to the $\hat{\theta}$ values. 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). 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, the ability of the M3PLMs to correctly identify speeded and nonspeeded examinees was assessed by computing the percentage of simulated speeded, nonspeeded, or total examinees whose modal group membership estimate identified the correct group. Finally, testlet variances were computed for each of the 3PLt models. Variances were computed separately for speeded and nonspeeded groups. Because LID was only simulated towards the end of the test, it was expected that the testlet variances would increase from the first to the last testlet. Also, because LID was not simulated for nonspeeded examinees, it was expected that the observed end-of-test testlet variances for speeded examinees would be greater than those for nonspeeded examinees.

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. For purposes of facilitating comparisons of results among models, the first 48 items were considered NS and the last 12 were considered SP, 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 biases in ability estimation. The largest bias, just 0.04, was for the 3PLt-2×4 model in the $\mathfrak{E}(\eta) = 0.90$ condition. Most of the models had overall biases that were no more than 0.02 in absolute value. In the $\mathfrak{E}(\eta) = 0.90$ condition, the NS ability biases for the models were all positive, but were also all quite similar. When $\mathfrak{E}(\eta) = 0.80$, ability was again consistently over-estimated for the NS group, but ability biases were noticeably higher for the 3PLt models than for either the 3PLM or the M3PLMs. In both conditions, for the SP group, biases were consistently negative for all models (i.e., ability was under-estimated), but were substantially more so for the 3PLt and 3PLM than for the M3PLMs.

RMSE data further demonstrated that the M3PLMs recovered underlying ability parameters best. For the 3PLt models and the 3PLM, RMSEs were large for all three groups, particularly for the SP group in the $\mathfrak{E}(\eta) = 0.80$ condition. In contrast, RMSEs for the M3PLM were small in all three groups in both conditions.

Results from the six 3PLt solutions varied, but all were consistently worse than the M3PLM solutions. Of the six models, biases and RMSEs for the 3PLt-2×4 appeared worst. No one 3PLt model emerged as consistently the best in terms of ability recovery. The 3PLM appeared to recover underlying ability parameters better than the 3PLt models, though differences were slight and its performance was clearly worse than that of the M3PLMs. Results from the three M3PLM solutions were virtually indistinguishable from one another.

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 contrast, the 3PLt models were unable to recover well the item difficulties for the speeded items, and also showed moderate positive bias for nonspeeded items when $\delta(\eta) = 0.80$. The 3PLM worked well at recovering nonspeeded items, but provided difficulties of the speeded items that were greatly over-estimated. RMSE data further demonstrated the poor recovery of item difficulties, particularly for the speeded items, from the 3PLt models and the 3PLM. Performance of the six 3PLt models was fairly similar. Again, results for the three M3PLM solutions were essentially identical.

Insert Table 3 About Here

Bias and RMSE results for item discrimination estimates (see Table 4) largely mirrored those for item difficulty and examinee ability. Biases and RMSEs were clearly smallest for the M3PLMs. That recovery of the a_i was so good is a bit surprising considering the M3PLMs constrain a_i to be equal for both the speeded and nonspeeded groups of examinees. Biases and RMSEs were, generally, appreciably larger for the 3PLt models and the 3PLM, particularly when $\delta(\eta) = 0.80$. The 3PLt-2×4 again appeared to provide the worst recovery. No other patterns were evident.

Insert Table 4 About Here

Correlations between generating and estimated θ parameters are provided in Table 5 for the different models. In the $\delta(\eta) = 0.90$ condition, correlations for the three groups were very similar for all models except the 3PLt-2×4, where they were slightly lower for the speeded and total groups. When $\delta(\eta) = 0.80$, the correlations for the 3PLt models were lower than for the other

models. In particular, the 3PLt-2×4 and the 3PLt-4×4 did not recover the rank ordering of ability well for speeded examinees.

Insert Table 5 About Here

In addition to examining the quality of parameter recovery, additional analyses were performed to determine whether the models fit the data and, for the M3PLMs and 3PLts, whether they accounted for the LID as expected. 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 (i.e., an average of 0.01) 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 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 $\mathfrak{E}(\eta)$, average Q_3 indices were 0.01 overall and 0.00 between pairs of end-of-test items. Despite the 3PLt including

examinee-specific parameters designed to absorb LID among sets of items, there was no difference between the 3PL and the 3PLt models in terms of average Q_3 values. All showed substantial positive LID among end-of-test items, with the amount of LID increasing as the items became closer to the end of the test. As expected, the amount of LID was higher in the $\mathfrak{E}(\eta)$ = 0.80 condition, where examinees became speeded earlier and saw their performance deteriorate more.

Model fit in the M3PLMs was further assessed by considering the classification accuracy of examinees. Table 7 shows the percentage of examinees who were correctly identified as speeded or nonspeeded by the model. In both speededness conditions, nonspeeded examinees were classified correctly with 96% accuracy, on average. Classification accuracy of speeded examinees, however, differed considerably by condition. Not surprisingly, it was easier to correctly classify speeded examinees in the $\mathfrak{E}(\eta)$ = 0.80 condition where their end-of-test item responses were more likely to be significantly contaminated. In the $\mathfrak{E}(\eta)$ = 0.90 condition, speeded examinees were correctly identified roughly 41% of the time, whereas in the $\mathfrak{E}(\eta)$ = 0.80 condition, they were correctly identified roughly 62% of the time.

Insert Table 7 About Here

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 γ values for speeded examinees than for nonspeeded examinees, particularly for testlets associated with end-of-test items. Testlet variances for the various 3PLt models are shown in Table 8. By and large, testlet variances increased from the first to the last testlet. For example, when $\mathfrak{E}(\eta)$ = 0.90, testlet variances for the 3PLt-4×4 increased from 0.10 (for the first testlet comprising items 1-44) to 1.12 (for the last testlet comprising items 57-60). However, there

were several exceptions. In the $\mathfrak{E}(\eta) = 0.90$ condition, $\sigma^2_{\gamma_1}$ and $\sigma^2_{\gamma_2}$ were often very similar or even identical. One reason for this may be that in this condition, speededness was not expected to begin until around item 55, and even then, the magnitude of performance degradation was expected to be minimal. Therefore, several of the supposedly speeded testlets contained items that were speeded for very few examinees. In addition, there were some models, particularly in the $\mathfrak{E}(\eta) = 0.80$ condition, where $\sigma^2_{\gamma_1}$ was substantially greater than $\sigma^2_{\gamma_2}$. In some of these models, in fact, $\sigma^2_{\gamma_1}$ was the largest of all the σ^2_{γ}.

Insert Table 8 About Here

Table 9 shows the differences between testlet variances for nonspeeded and speeded examinees under the two conditions. Differences were computed as the nonspeeded variance minus the speeded variance. Testlets associated with end-of-test items were expected to show LID for the speeded group. Therefore, on these testlets, σ^2_{γ} should be higher for the speeded examinees, resulting in negative values in Table 9. In fact, this was precisely the observed trend. Differences in testlet variances were positive for the first testlet, but were negative for most testlets thereafter. Furthermore, the differences in variances for testlet *k* were always greater than or equal to those for testlet *k*+1. This suggests that the speededness effect became more pronounced as examinees progressed through the test, and the γ parameters in the 3PLt attempted to compensate for it.

Insert Table 9 About Here

Discussion

This study compared item and ability parameter recovery for two models for addressing LID due to test speededness, the 3PLt and the M3PLM, with that of the traditional 3PLM that

assumes all items are locally independent. The results of the study showed quite clearly that the M3PLM recovered underlying parameters better than the other two models. Regardless of the amount of simulated speededness, the M3PLM had low biases and RMSEs for both item parameters and examinee trait parameters, even for the end-of-test items and speeded examinees. Varying the number of end-of-test items on which ordinal constraints were placed resulted in only trivial differences.

In contrast, the 3PLt and 3PL produced item parameters that were positively biased for the speeded items and examinee trait parameters that were negatively biased for speeded examinees. The amount of bias increased as the amount of speededness increased. RMSEs for speeded groups for both item and examinee parameters were substantially larger than the corresponding values for the M3PLM. The 3PLt appeared slightly better than the 3PLM at recovering underlying item parameters, though recovery of trait parameters was comparable in the two models, if not slightly better in the 3PLM.. No one 3PLt model emerged as best, though the 3PLt-2×4 was often among the worst at recovering parameters.

The similarity of the 3PLM and 3PLt is a bit surprising considering the 3PLt has worked well previously at handling different types of LID and the 3PLM is unequipped to deal with LID of any form. There are a few reasons that the anticipated differences might not have been realized here. First, although the item and ability parameter estimates for both models were equated to the metric of the generating parameters, the equatings for the 3PLt did not utilize information about the differences between estimated and true testlet means when estimating the transformation coefficients. This information was not included because true testlet means did not exist, since the data were not generated by the testlet model. However, failing to include information on the locations of the testlets could have led to inaccurate estimates of the equating

coefficients.

A second factor that may have contributed to the unexpected findings is that testlet variances for all testlets were estimated. Because the items in the first testlet we all assumed to be nonspeeded, the testlet variance for this testlet could have been constrained to equal zero. In effect, this would have forced a 3PLM onto the estimation of these items. Instead, all variance parameters were estimated and, in several cases, variances for the nonspeeded testlet were substantially larger than for the speeded testlets. Had nonspeeded variances been fixed at zero, it is possible that the other variances would have changed accordingly to better compensate for the modeled LID.

Finally, it is possible that the 3PLt is not well suited for handling LID due to test speededness. Test speededness manifests itself in very different ways than does, for example, a passage effect. When passage effects are present, all examinees seeing those items are subject to the effect. Certainly the strength of the effect will vary across individuals, but it is assumed that the impact on the items is constant for a particular individual. For example, if a particular passage is interesting and easy to understand for examinee j, all the corresponding items will be made easier by a fixed amount (equal to γ_j), whereas if examinee k finds the passage difficult to understand, all items will be made harder by γ_k. With test speededness, it is neither true that all examinees are subject to the effect, nor that it is a constant effect for those who do experience it. On the tests simulated here, 75% of the examinees were able to answer all items, hence all items were simulated as locally independent. Among the 25% for whom there was a speededness effect, the effect was simulated to become stronger and stronger throughout the exam. Therefore, in the 3PLt-1×8 condition, for example, given (a_i, b_i, c_i) parameters of (1, 0, 0.2), an examinee with $\theta = 1.0$, $\lambda = 0.88$ and $\eta = 10$ would correctly answer items 1-52 with probability

0.78, item 53 with probability 0.77, and item 60 with probability 0.36. Because the 3PLt assigns only a single γ per testlet, for many examinees it may not be possible to find a value that adequately describes the speededness effect across that set of items.

Based on results from this study, the M3PLM appears to be an excellent choice for estimating item and examinee parameters for speeded tests. However, the legality of using the M3PLM for trait estimation remains questionable. One possibility is to use the M3PLM to find good, stable estimates of the item parameters, and then use these item parameters to estimate the ability level for all examinees (see Wollack et al., 2003). However, doing this would undoubtedly reintroduce negative bias into the ability estimates for speeded examinees. It remains to be seen whether this estimation technique would produce levels of bias that are lower than those from the 3PLM or 3PLt.

Another alternative is the hybrid model (Yamamoto & Everson, 1997). Like the M3PLM, the hybrid model is a latent class model designed to account for test speededness, but, like the 3PLt, is parameterized so that only a single set of item parameters exists for all examinees. Future study should involve comparing the parameter recovery of the hybrid model to that of the M3PLM.

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	-0.04	0.01	0.26	0.28	0.26
M3PLM-8	0.03	-0.04	0.01	0.26	0.29	0.27
M3PLM-16	0.03	-0.03	0.02	0.26	0.28	0.27
3PLt-1×16	0.02	-0.11	-0.01	0.53	0.60	0.55
3PLt-2×8	0.01	-0.09	-0.01	0.53	0.57	0.54
3PLt-4×4	0.01	-0.10	-0.01	0.53	0.58	0.54
3PLt-1×8	0.03	-0.14	-0.01	0.58	0.67	0.61
3PLt-2×4	0.04	-0.15	0.04	0.59	0.69	0.62
3PLt-1×4	0.03	-0.14	-0.01	0.61	0.70	0.64
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	0.26	0.30	0.27
M3PLM-8	0.03	-0.03	0.02	0.26	0.29	0.27
M3PLM-16	0.04	-0.02	0.02	0.26	0.29	0.27
3PLt-1×16	0.08	-0.17	0.01	0.56	0.70	0.59
3PLt-2×8	0.09	-0.20	0.02	0.56	0.72	0.61
3PLt-4×4	0.11	-0.28	0.01	0.58	0.80	0.64
3PLt-1×8	0.07	-0.15	0.02	0.59	0.71	0.62
3PLt-2×4	0.12	-0.30	0.02	0.66	0.89	0.72
3PLt-1×4	0.07	-0.15	0.01	0.60	0.72	0.63
3PLM	0.02	-0.19	-0.03	0.50	0.65	0.54

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	0.03	0.04	0.01	0.01	0.01
M3PLM-8	0.04	0.03	0.04	0.01	0.01	0.01
M3PLM-16	0.04	0.03	0.04	0.01	0.01	0.01
3PLt-1×16	0.03	0.12	0.05	0.05	0.16	0.08
3PLt-2×8	0.03	0.13	0.05	0.05	0.17	0.09
3PLt-4×4	0.03	0.12	0.05	0.05	0.16	0.08
3PLt-1×8	0.04	0.12	0.05	0.05	0.15	0.08
3PLt-2×4	0.04	0.11	0.05	0.05	0.14	0.08
3PLt-1×4	0.03	0.10	0.04	0.05	0.13	0.07
3PLM	0.01	0.17	0.05	0.07	0.24	0.12
$\mathcal{E}(\eta) = 0.80$						
M3PLM-4	0.06	0.01	0.05	0.01	0.01	0.01
M3PLM-8	0.05	0.01	0.05	0.01	0.01	0.01
M3PLM-16	0.05	0.01	0.05	0.01	0.01	0.01
3PLt-1×16	0.08	0.24	0.11	0.09	0.25	0.14
3PLt-2×8	0.09	0.23	0.11	0.10	0.23	0.14
3PLt-4×4	0.09	0.21	0.12	0.10	0.22	0.14
3PLt-1×8	0.07	0.23	0.10	0.09	0.24	0.13
3PLt-2×4	0.09	0.21	0.12	0.11	0.21	0.14
3PLt-1×4	0.06	0.26	0.10	0.08	0.27	0.14
3PLM	-0.01	0.35	0.06	0.08	0.36	0.17

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	0.06	0.04	0.01	0.02	0.01
M3PLM-8	0.04	0.06	0.04	0.01	0.02	0.01
M3PLM-16	0.03	0.07	0.04	0.01	0.02	0.01
3PLt-1×16	0.05	0.01	0.04	0.09	0.09	0.09
3PLt-2×8	0.05	0.02	0.04	0.09	0.10	0.09
3PLt-4×4	0.04	0.04	0.04	0.09	0.10	0.09
3PLt-1×8	0.03	0.09	0.05	0.08	0.14	0.10
3PLt-2×4	0.02	0.13	0.05	0.08	0.18	0.11
3PLt-1×4	0.04	0.05	0.04	0.08	0.17	0.10
3PLM	0.06	-0.01	0.04	0.12	0.10	0.11
$\mathcal{E}(\eta) = 0.80$						
M3PLM-4	0.04	0.06	0.05	0.01	0.02	0.01
M3PLM-8	0.04	0.06	0.04	0.01	0.02	0.01
M3PLM-16	0.04	0.06	0.04	0.01	0.02	0.01
3PLt-1×16	-0.02	0.15	0.01	0.08	0.21	0.12
3PLt-2×8	-0.04	0.14	0.00	0.10	0.19	0.12
3PLt-4×4	-0.09	0.27	-0.02	0.13	0.30	0.18
3PLt-1×8	0.03	0.11	0.05	0.09	0.19	0.12
3PLt-2×4	-0.13	0.25	-0.05	0.15	0.43	0.23
3PLt-1×4	0.09	0.19	0.11	0.13	0.28	0.17
3PLM	0.02	0.27	0.07	0.11	0.37	0.19

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	0.92	0.93
3PLt-2×8	0.94	0.93	0.93
3PLt-4×4	0.94	0.92	0.93
3PLt-1×8	0.93	0.91	0.91
3PLt-2×4	0.92	0.89	0.90
3PLt-1×4	0.93	0.91	0.91
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	0.89	0.90
3PLt-2×8	0.92	0.86	0.87
3PLt-4×4	0.91	0.81	0.83
3PLt-1×8	0.93	0.90	0.90
3PLt-2×4	0.88	0.75	0.76
3PLt-1×4	0.93	0.89	0.90
3PLM	0.94	0.90	0.91

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	0.02	0.03	0.06	0.09
3PLt-2×8	0.01	0.02	0.04	0.06	0.09
3PLt-4×4	0.01	0.02	0.03	0.06	0.09
3PLt-1×8	0.01	0.02	0.03	0.05	0.08
3PLt-2×4	0.01	0.01	0.02	0.05	0.08
3PLt-1×4	0.01	0.02	0.04	0.06	0.08
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	0.01	0.09	0.10	0.11	0.11
3PLt-2×8	0.01	0.08	0.10	0.11	0.11
3PLt-4×4	0.01	0.07	0.09	0.10	0.10
3PLt-1×8	0.01	0.08	0.10	0.11	0.11
3PLt-2×4	0.02	0.06	0.08	0.09	0.09
3PLt-1×4	0.01	0.07	0.09	0.10	0.10
3PLM	0.00	0.07	0.09	0.10	0.10

Table 7. Classification Accuracy of the M3PLMs

	Classification Accuracy		
	NS	SP	Total
$\xi(\eta) = 0.90$			
M3PLM-4	96%	40%	82%
M3PLM-8	96%	41%	82%
M3PLM-16	95%	42%	82%
$\xi(\eta) = 0.80$			
M3PLM-4	96%	61%	88%
M3PLM-8	96%	62%	88%
M3PLM-16	96%	62%	88%

Table 8. Testlet Variances

	Variances				
	$\sigma^2_{\gamma_1}$	$\sigma^2_{\gamma_2}$	$\sigma^2_{\gamma_3}$	$\sigma^2_{\gamma_4}$	$\sigma^2_{\gamma_5}$
$\mathcal{E}(\eta) = 0.90$					
3PLt-1×16	0.10	0.10			
3PLt-2×8	0.08	0.08	0.59		
3PLt-4×4	0.10	0.13	0.15	0.23	1.12
3PLt-1×8	0.29	0.33			
3PLt-2×4	0.36	0.13	0.72		
3PLt-1×4	0.40	0.70			
$\mathcal{E}(\eta) = 0.80$					
3PLt-1×16	0.30	0.39			
3PLt-2×8	0.40	0.08	0.69		
3PLt-4×4	0.59	0.19	0.15	0.32	0.37
3PLt-1×8	0.31	0.78			
3PLt-2×4	0.93	0.15	0.18		
3PLt-1×4	0.25	0.66			

Table 9. Differences Between Testlet Variances
for nonspeeded and Speeded Examinees

	Differences in Variances (NS − SP)				
	$\sigma^2_{\gamma_1}$	$\sigma^2_{\gamma_2}$	$\sigma^2_{\gamma_3}$	$\sigma^2_{\gamma_4}$	$\sigma^2_{\gamma_5}$
$\mathcal{E}(\eta) = 0.90$					
3PLt-1×16	0.07	-0.08			
3PLt-2×8	0.05	0.01	-0.48		
3PLt-4×4	0.06	0.01	0.01	-0.07	-0.07
3PLt-1×8	0.18	-0.29			
3PLt-2×4	0.21	-0.03	-0.48		
3PLt-1×4	0.16	-0.47			
$\mathcal{E}(\eta) = 0.80$					
3PLt-1×16	0.25	-0.50			
3PLt-2×8	0.32	-0.03	-0.70		
3PLt-4×4	0.47	0.03	-0.04	-0.20	-0.20
3PLt-1×8	0.17	-0.78			
3PLt-2×4	0.49	-0.10	-0.13		
3PLt-1×4	0.07	-0.47			