A Model for Simulating Speeded Test Data

James A. Wollack
University of Wisconsin–Madison

Allan S. Cohen
University of Georgia

April 13, 2004

Portions of this paper were presented at the annual meeting of the American Educational Research Association, San Diego, CA.

RUNNING HEAD: Speededness Generating Model

A Model for Simulating Speeded Test Data

Tests consisting of items that violate the item response theory (IRT) assumption of local independence can cause serious problems for test developers. The inclusion of items with local item dependence (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.

Yen (1993) identified ten causes of local dependence. One of the most prevalent causes in educational testing is test speededness. 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. 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, and Wells, 2003).

In the past few years, several models have been developed that account for local dependencies due to speededness effects. Yamamoto and Everson (1997) proposed 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. Yamamoto

and Everson's hybrid model identifies K + 1 latent groups of examinees, one for whom the model, e.g., the 3PL, is appropriate for all items, and K latent groups whose patterns better approximate random guessing on the last k = 1, 2, 3, . . . , K items of the test. Bolt, Cohen, and Wollack (2002) proposed a mixture Rasch model (MRM) which assumes that two latent classes of examinees exist, and that the Rasch model is appropriate for both classes. For items early in the test, the MRM item difficulty parameters are constrained to be equal in the two classes; however, for end-of-test items, the item difficulty parameters are constrained to be larger (i.e., harder) in one of the classes. The Bolt et al. mixture method was extended to the 2-parameter logistic model (2PL) and the 3PL by Bolt, Mroch, and Kim (2003). Wollack, Wells, and Cohen (2003) developed a PCM extension of the mixture model approach, thereby modeling LID due to both speededness and passage effects.

In addition, the 3PL testlet model (3PL-t; Bradlow, Wainer, & Wang, 1999; Wainer, Bradlow, & Du, 2000) has been proposed to explicitly model the systematic nuisance variation that commonly exists among items within a testlet. The 3PL-t is modeled by the inclusion of a random effects, testlet- and examinee-specific (parameter which is subtracted from the 3PL item difficulty. Wainer et al. (2000) and Li & Cohen (2003) found the 3PL-t to work better than other available models for accounting for LID. Because the 3PL-t does not specifically model any particular type of LID, only testlet-specific LID, it is conceivably that it could be used to account for LID due to speededness, and quite possibly, for LID due to both speededness and passage effects.

Because models of test speededness are relatively new, few studies exist comparing different test speededness models. The studies that do exist (Bolt et al., 2003; Cohen, Wollack,

Bolt, & Mroch, 2002) describe differences in the types of examinees identified as speeded and compare the similarity of item parameter estimates under both models for items when administered under speeded or assumed nonspeeded conditions. However, no simulation studies have been conducted to compare the classification accuracy of these models, in large part because realistic speededness simulators do not exist. In this paper, we develop and validate a model for generating speeded test data. Simulated datasets from this model can then be used to make comparisons among existing item response models for dealing with test speededness.

Developing a Model for Simulating Speeded Test Data

To make valid comparisons among models, it is necessary to generate data that are as realistic as possible. The two general models for describing and estimating model parameters in speeded data, the hybrid model and the mixture model, while useful, are both oversimplifications of test speededness. This is particularly true if they were to be used for data generation. Generating data under the hybrid model would mean that examinees respond according to the model up to a point, after which their responses are completely random. That is,

$$P_i^*(q_j) = \begin{cases} c_i + (1-c_i)P_i(q_j); & i < k_j \\ c_i & ; \; i \geq k_j \end{cases}$$

where $P_i^*(2_j)$ is the probability of a speeded examinee with ability 2_j answering item i (i = 1, . . . , n) correctly, $P_i(2_j)$ is the standard two parameter logistic model, $P_i(q_j) = \frac{1}{1+e^{-a_i(q_j - b_i)}}$, such that $"_i$ and $\$_i$ are the discrimination and difficulty parameters, respectively, for item i, and c_i is a random guessing parameter for item i, fixed to equal to the reciprocal of the number of alternatives, and k_j is the item number of the first speeded item for examinee j. Speededness is unlikely so straightforward, as no doubt some students are hurried, but do not resort to random guessing.

Generating data under the mixture model, meanwhile, assumes that there is some point in the test before which all examinees fit a common model, and after which all examinees in the speeded class find the test more difficult, as shown below:

$$P_i^*(q_j) = \begin{cases} [1+\exp(-(q_j - b_i))]^{-1}; & i < k \\ [1+\exp(-(q_j - b_{i,S}))]^{-1}; & i \geq k \end{cases}$$

where $b_{i,S}$ are the Rasch item difficulties for the examinees in the latent speeded class (see Bolt et al., 2002). Though this model has worked well at identifying test speededness, from a simulation perspective, it is likely overly simplistic. In fact, speededness is quite complex. Different examinees become speeded at different points in the test. Some examinees will devote the necessary time to each item they attempt, guessing on the others, while some examinees will manage their time differently to make sure they have some time to read and attempt every question.

We propose the following model for simulating realistic speeded data:

$$P_i^* = c_i + (1-c_i)\left\{P_i(q_j)\cdot \min\left(1, \left[1-\left(\tfrac{i}{n}-h_j\right)\right]\right)^{l_j}\right\};$$

where h_j ($0 \leq h_j \leq 1$), and l_j ($l_j \geq 0$) are the speededness point parameter and speededness rate parameter of examinee j, respectively, min [x, y] is the smaller of the two values x and y, and c_i, q_j, and $P_i(q_j)$ are as previously defined.

This model is appealing in that it allows for different examinees to become speeded at different points in the test, and also allows them to have different rates of decline. In this model, the examinee is speeded on all items beyond the $(n\times h)^{th}$ item (i.e., when $h \geq i/n$), and the examinee's probability of answering the item correctly is reduced by a some fraction, the size of which is determined by two parameters. The speededness point parameter, h_j, identifies the point in the test, expressed as a proportion of the items completed, at which an examinee first

experiences an effect due to speed. As an example, an $\theta_j = .75$ indicates that examinee *j* becomes speeded three-quarters of the way through the test. The farther beyond the initial speededness point an examinee progresses, the larger $\left(\frac{i}{n} - b_j\right)$ becomes causing a greater reduction to $P_i^*(\theta_j)$. The second parameter influencing $P_i^*(\theta_j)$ in this model is the speededness rate parameter, β_j. Once an examinee passes the speededness point, $\left(\frac{i}{n} - b_j\right)$ is raised to the power β_j, which serves to control the speed at which $P_i^*(\theta_j)$ decreases. At its most extreme, when $\left[1 - \left(\frac{i}{n} - b\right)\right]^{\beta_j} = 0$, this model reduces to the hybrid model, and speededness is modeled by random guessing. When $\theta > i/n$, examinee *j*'s performance is not (yet) speeded and $P_i^*(\theta_j)$ reduces to the 3PL. Examinees with θ values of 1.0 or β values of 0.0 are not speeded for any items.

Simulating Speeded Data

Item and examinee parameters estimated using MULTILOG (Thissen, 1991) from data on nearly 4,445 examinees on a 60-item college-level reading comprehension test were used as generating parameters for the hybrid model, MRM, and our speededness generating model (SGM). This dataset was believe to contain inadvertent speededness, due to the imposition of time limits. The reading comprehension test consisted of 11 reading passages, each followed by between 5 and 7 items. Item parameters were estimated using both the three parameter model with the c parameter fixed at .2 (3PL-c) and the MRM. To estimate the difficulties in the MRM, the seven items associated with the last passage were constrained to be harder for the speeded group than for the nonspeeded group.

Item parameter estimates from the 3PL-c were used as generating parameters to simulate item responses to nonspeeded items for the hybrid model and SGM. Speeded items with the

hybrid model were simulated as random responses which would be correct 20% of the time and incorrect 80% of the time. Two different degrees of speededness severity were simulated, represented by the generating distribution for the θ parameters. θ parameters for speeded examinees were generated according to a beta (9, 2) distribution to simulate moderately high speededness, and from a beta (20,2) distribution to simulate moderately low speededness. To understand better the impact of these distributions on speededness severity, Table 1 reports the expected percentage of examinees who will be speeded at different points in a 60-item test. A lognormal (3.912,1) distribution was used for β parameters. The estimated MRM difficulties for the speeded and nonspeeded classes were used as generating parameters for simulating data in the MRM.

Insert Table 1 About Here

For all three models, 5 datasets of 2,000 examinees each were simulated for a 60-item test. Estimates of α from a random 2,000 of the 4,445 examinees completing the reading comprehension test were treated as generating ability parameters for all datasets in all three models. Although the same α vector and item parameters were used to simulate all datasets, θ_j and β_j were generated separately for each replication. Within each dataset, the first 500 examinees were simulated as speeded, and the remaining 1,500 as nonspeeded. To facilitate comparisons with the MRM, for the hybrid model, one seventh of the examinees (either 71 or 72 examinees) were simulated as being speeded on the last seven items, one seventh on the last six items, one seventh on the last five items, and so forth, with the final one seventh of the speeded examinees being speeded on only the last item.

Methods

Simulated data from three models–hybrid model, MRM, and SGM–will be compared, with particular interest being paid to the extent to which the simulated data match our expectations for response behavior when examinees are speeded. It is very difficult to have a firm handle on what speeded data should look like in practice, as speededness is a latent, unobservable trait. Our basic premise, however, is that different examinees experience speededness differently. Our belief is that, in practice, this should manifest itself in two important ways:

1. Because it is known that examinees work at different rates, examinees should first experience the effects of speededness at different points in the test.
2. Strategies for dealing with time constraints differ across examinees. Some examinees will answer all the items they can in the allotted time, and then guess randomly on however many items remain, while other examinees will, at some point in the test, devote less than the optimal amount of time to each remaining item, in order to give themselves an opportunity at all items. Therefore, some examinees will resort to guessing shortly after becoming speeded, while others will see gradually diminished performance on end-of-test items.

To investigate how well the expectation that examinees become speeded at different points is satisfied under the various speededness simulation models, we compared the models with respect to the differences in item-level percentage correct scores between simulated speeded and nonspeeded examinees and the distribution of locations of the first speeded item for speeded examinees. Expectations about the different rates of deterioration due to speededness were studied by comparing the models with respect to the difference between observed percentage correct scores on the first speeded item and the expected value, had the examinees not been

speeded. The rate of diminished performance was also studied by examining, for all models, $P_i^*(2_j)$ for examinees with 2 values of !2, !1, 0, 1, and 2, on the last eight items of the reading comprehension test.

Results

Tables 2 through 5 show, for the hybrid model, MRM, and SGM with 0 ~ beta(20,2) (hereafter referred to as SGM(20,2)) and SGM with 0 ~ beta(9,2) (hereafter referred to as SGM(9,2), respectively, the percentage correct scores for every third item, beginning with item 1, through item 30, and all items afterwards, for the nonspeeded, speeded, and total groups, as well as the difference in percentages between the speeded and nonspeeded groups. Differences in percentage correct scores between speeded and nonspeeded examinees were near zero for all but the items at the end of the test. Speededness occurred soonest in the SGM(9,2) model (Table 5), producing noticeable differences beginning with item 41. That speededness occurred this early in this model is not surprising; 41/60 = .683 corresponds to the 12th percentile in the beta(9,2) distribution, meaning that 12% of the examinees were expected to have encountered their first speeded item before item 41. Differences didn't become consistently double-digits until item 49, the mean of the beta(9,2) distribution. Though there was some fluctuation due to differences in item difficulty, the differences between the speeded and nonspeeded groups tended to increase gradually as the test progressed. The items at the very end of the test often produced rather large differences in performance between the two groups.

Insert Tables 2 - 5 About Here

The SGM(20,2) produced similar results (Table 4), although the location where speededness first became apparent changed. In this model, differences between speeded and nonspeeded examinees became noticeable around item 50, which curiously also corresponds to the 12th percentile of the generating θ distribution. Differences gradually increased; differences on the items at the end of the test were large, though not as large as when a beta(9,2) distribution was used for θ.

The hybrid model and MRM both showed first signs of speededness at item 54 (see Tables 2 and 3). This is not unexpected, as speededness was only simulated with these models for items 54 through 60. However, as these models require the specification of the point at which speededness first occurs, this effect is unavoidable. Because we simulated one-seventh of the speeded examinees becoming speeded on each of items 54 through 60, the differences between speeded and nonspeeded examinees increase gradually in the hybrid model. Except for the fact that SGM (20,2) showed signs of speededness earlier than the hybrid model, the differences in percentage correct look strikingly similar between those two models. The MRM, on the other hand, produces a very different picture of test speededness. In the MRM, items almost immediately begin performing at or below the chance level for the speeded group, resulting in very large differences between the speeded and nonspeeded groups. With the MRM, the effect of speededness appears uniform across all speeded items.

Although the hybrid model and SGM(20,2) produced similar percentage correct scores aggregated across all examinees, a very different picture is revealed by inspecting the distribution of items on which speededness was first encountered. Table 6 provides the percentage of simulated speeded examinees who became speeded at different points in the test. Note that examinees who became speeded prior to item 31 are combined into a single category. Only the SGM(9,2) condition provided any examinees in the “Items 1 - 30” category. The

earliest in the test that an examinee became speeded in the SGM(9,2) condition was at item 24.

Insert Table 6 About Here

From Table 6, it is clear that SGM(9,2) and SGM(20,2) produce the expected effect of having examinees become speeded at different points in the test. The particular shape and parameters of the distribution used for θ control how early they become speeded and the point in the test at which most of the examinees will experience speededness. The hybrid model and MRM, on the other hand, produce rather unrealistic speededness patterns. In the hybrid model, exactly one-seventh of the examinees become speeded at each of the last seven items. In the MRM, all of the examinees experience speededness at precisely the same time. Both the hybrid model and MRM, as generating models, do not provide examinees with enough flexibility to become speeded at different times.

To examine the rate at which performance drops off for speeded examinees, we examined the difference between expected and observed percentage correct scores on the first speeded item. These percentages, along with the difference between the two, are given in Table 7 for the hybrid model and MRM, Table 8 for the SGM(20,2), and Table 9 for the SGM(9,2). Because the hybrid model, SGM(20,2), and SGM(9,2) all model the probability of correct response for nonspeeded examinees with the 3PL, the expected percentage correct scores for these models were computed as the sum of 3PL probabilities of correct response across all examinees becoming speeded on item *i*. By summing across only those examinees becoming speeded on item *i*, we account for any differences in mean θ scores among the groups of examinees becoming speeded on different items. For the SGM(20,2) and SGM(9,2), data are only provided

for those items on which at least five examinees first experienced speededness. For the MRM, because all speeded examinees became speeded on item 54, the expected percentage correct score was estimated as the sum of the Rasch-based $P(2|\beta_{54,NS})$ across all examinees in the speeded group, where $\beta_{54,S}$ is the Rasch item difficulty for item 54 for the nonspeeded group.

Insert Tables 7 - 9 About Here

We know from Tables 2 and 3 that before speededness is simulated in the hybrid model and MRM, examinees in the speeded and nonspeeded groups perform very similarly. From Table 7, we see that as soon as examinees become speeded, their performance immediately drops off dramatically. In the hybrid model, the differences between observed and expected performance range from .12 to .48, with those differences being explained entirely by differences in item difficulty. Immediately when an examinee becomes speeded in the hybrid model, the probability of that examinee getting an item correct becomes .20, the reciprocal of the number of item alternatives. The MRM also exhibits the dramatic change in observed versus expected performance. Whereas the fact that the observed proportion correct for the speeded group is .20 in the hybrid model is a byproduct of the model, in the MRM, it is somewhat of a coincidence. The probability for the nonspeeded group, in theory, could be substantially higher, or it could be as low as zero.

Data from the SGM(20,2) and SGM(9,2), however, paint a very different picture. With the exception of differences of .28 and .57 on items 29 and 30 for SGM(9,2), both of which were based on fewer than 10 people, the differences between expected and observed performance never varied by more than .23, were less than .10 in 58 percent of the cases. This suggests that,

on average, performance does not drop off as quickly and dramatically in the SGM(20,2) and SGM(9,2) as it does in the hybrid model and MRM.

Tables 7 - 9 provide insight into how examinees perform on the first speeded item. To gain perspective on the rate of decline beyond the first speeded item, Table 10 provides the probability of examinees with different θ values selecting the correct response to each of the last 8 items on the test. Because the difficulties of the items differ, to facilitate comparisons, the 3PL probabilities of correct response are provided under the assumption of no speededness. Note that in the hybrid model and MRM, speededness was only simulated on the last 7 items. Also note that, for purposes of understanding the rate of change, ω was fixed to be either .90 or .80 for the SGM model, while τ was held constant at 3.912.

Insert Table 10 About Here

As was observed in Table 2, with the hybrid model, group-level performance appears to drop off slowly. Recall that one seventh of the examinees are assumed to become speeded on each of items 54 through 60, so item 56, for example, has four-sevenths of the examinees answering based on the 3PL, while three-sevenths answer with probability 0.2. On the last item, all examinees are speeded, so $P_i^*(\theta_j) = 0.2$ for all θ levels.

For the MRM, performance appears to drop off very quickly, particularly for examinees with $\theta < 2$. In fact, for examinees with $\theta \# 0$, MRM $P_i^*(\theta_j)$ values are very often less than chance, and in some cases are very nearly zero. Unless for such examinees being speeded means that they are likely to omit responses, observing probabilities significantly less than chance would appear inconsistent with a speededness hypothesis that, at its worst, would predict that these examinees would randomly fill in answers.

SGM (θ = .9) and SGM (θ = .8) both show gradual rates of decline. Under both models, even for examinees with ζ = !2, on the last item on the test, $P_i^*(\zeta_j)$ remains above 0.2. The very good speeded student (i.e., ζ = 2), even in the SGM (θ = .8) condition where examinees become speeded after item 48, still has a higher $P_i^*(\zeta_j)$ than a nonspeeded student with ζ = !1, and, until item 57, has a higher $P_i^*(\zeta_j)$ than a nonspeeded student with ζ = 0. In the SGM (θ = .9) condition where examinees become speeded beginning with item 55, examinees with ζ = 2 have a higher $P_i^*(\zeta_j)$ than a nonspeeded examinee with ζ = 1 for all but items 58 and 60.

All of the SGM data to this point have concentrated on the performance of this model for different values of θ. To a certain extent, θ affects both the amount and rate of speededness. Obviously, θ directly controls the location at which speededness first occurs. However, θ indirectly affects the rate at which performance decreases because, for any given value of β, the further θ is from i/n, the bigger the reduction on $P_i^*(\zeta_j)$. A more direct way to affect the rate at which speededness impacts $P_i^*(\zeta_j)$ is by altering the β value. To study the impact of different values of both θ and β, we examined how $P_i^*(\zeta_j)$ changes over the last 10 items on a 60-item test, as a function of different values of ζ, for θ values of .8 and .9, and β values of 2.0, 4.0, and 8.0. So as to best show the impact of these variables on rate of decline, item parameters for the last 10 items were fixed at α = 1.0, γ = 0.0, and c = 0.2 for all items. These values are given in Table 11.

Insert Table 11 About Here

The ultimate differences between the two θ values and three β are, perhaps, best seen by inspecting the column for item 60, although examining the results for several different items gives a better picture of how the final results came to pass. For ζ = !2, and to a certain extent

for $\theta = -1$, the percentages do not differ much. However, for $\theta \geq 0$, there are some clear differences. The β parameter appears to influence $P_i^*(\theta_j)$ more than the α parameter. As an example, for examinees with $\theta = 1$, on item 60, the differences between $P_i^*(\theta_j)$ for $\alpha = .8$ and $\alpha = .9$ were .10, .14, and .15 for $\beta = 2.0$, 4.0, and 8.0, respectively. Differences between $P_i^*(\theta_j)$ for $\beta = 2.0$ and $\beta = 8.0$ were .22 and .27 for $\alpha = .9$ and .8, respectively. Larger differences were observed for $\theta = 2$.

Conclusion

SGM is a very flexible model for simulating speeded data. The analyses in this paper show that SGM allows for examinees to become speeded at different points, and for their drop in performance due to speededness to vary from gradual to very steep. The point at which a test becomes speeded and the rate of diminished performance can be altered through the particular distribution used for α and β. In this paper, they were modeled with the beta and lognormal distributions, respectively, though other distributions could certainly be used.

The hybrid model and MRM have both been shown to be useful models for identifying speeded examinees; however this paper demonstrates that neither model is appropriate for simulating speeded data. The MRM is overly simplistic, in that it assumes that all speeded examinees become speeded at the same point in the test and experience the same rate of decline. It is worth mentioning that the assumption that all examinees become speeded at the same point is unique to the situation where the MRM is used to simulate speeded data. When estimating parameters using the MRM, it is necessary to identify a set of items at the end of the test on which speeded examinees perform less well than nonspeeded examinees, but it is not necessary to assume that all examinees become speeded at a particular, common point. In fact, Wollack et

al. (2003) showed that using a MRM constrained so that the last six items were more difficult for examinees in the speeded group, a table of the differences in proportion correct scores for the speeded and nonspeeded classes suggested that the test became speeded for some examinees well before the last six items.

The hybrid model does allow for examinees to become speeded at different points in the exam. In this study, to maximize the comparison with the MRM, it was decided to have one-seventh of the examinees become speeded on each of the last seven items. Admittedly, the choice of a uniform distribution over the last seven items was arbitrary. We could have selected a beta(20,2) or beta(9,2) distribution, for example, to determine how many examinees would become speeded at different points during the test. However, because the hybrid model doesn't allow for an examinee gradually becoming more affected by speededness, it is unlikely that this would have improved its performance in this study. From the perspective of producing realistic simulated data, the fact that the hybrid model allows for different speededness points does not offset the fact that, in the hybrid model, speeded examinees respond at random to all speeded items.

Developing a speededness simulator that produces realistic results is critical if the various speededness models are to be compared in any meaningful way. The next step in this project, is to combine the SGM model with a model for simulating sets of correlated item responses, as might be observed with items associated with a common reading passage. Datasets from the combined model would, therefore, have up to two different sources of LID: speededness and passage dependence. Several different models for dealing with one or both sources of LID will be applied to simulated datasets to begin to better understand the relative strengths and weaknesses of the various models for handling locally dependent data structures.

References

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application 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.

Cohen, A. S., Wollack, J. A., Bolt, D. M., Mroch, A. A. (April, 2002). *A mixture Rasch model analysis of test speededness*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

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.

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. & 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.

Sireci, S. G., Thissen, D. & Wainer, H. (1991). On the reliability of testlet-based tests.

Journal of Educational Measurement, 28, 237-247.

Thissen, D. (1991). *MULTILOG: Multiple, categorical item analysis and test scoring using item response theory* [Computer program]. Chicago, IL: Scientific Software, Inc.

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., & Wells, C. S. (2003). The effects of test speededness on score scale stability. *Journal of Educational Measurement, 40*, 307-330.

Wollack, J. A., Wells, C. S., & Cohen, A. S. (April, 2003). *A comparison of item- and testlet-level scoring on scale stability in the presence of test speededness*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

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

Table 1

Percentages of Examinees Expected to be Speeded

	1- 3 items	4 - 6 items	7 - 9	10 - 12	> 12 items
beta (20, 2)	.283	.352	.210	.097	.058
beta (9, 2)	.086	.178	.192	.168	.376

Table 2

Percentage Correct Scores for Hybrid Model

Item #	Total Group	Nonspeeded Group	Speeded Group	Difference
1	.71	.71	.72	-.01
4	.63	.63	.65	-.02
7	.65	.65	.66	-.01
10	.64	.64	.63	.01
13	.62	.62	.61	.01
16	.77	.76	.78	-.02
19	.58	.58	.58	.00
22	.59	.59	.58	.01
25	.64	.64	.65	-.01
28	.68	.68	.68	.00
31	.58	.58	.59	-.01
32	.76	.76	.76	.00
33	.61	.60	.61	-.01
34	.62	.61	.62	-.01
35	.68	.68	.68	.00
36	.62	.61	.62	-.01
37	.47	.47	.46	.01
38	.67	.67	.68	-.01
39	.43	.43	.43	.00
40	.53	.52	.53	-.01
41	.64	.64	.64	.00
42	.59	.58	.60	-.02
43	.61	.61	.62	-.01
44	.65	.65	.63	.02
45	.69	.69	.68	.01
46	.56	.56	.56	.00
47	.59	.59	.58	.01
48	.56	.56	.57	-.01
49	.66	.65	.67	-.02
50	.68	.67	.68	-.01
51	.60	.60	.60	.00
52	.63	.63	.64	-.01
53	.54	.54	.54	.00
54	.61	.63	.55	.08
55	.38	.40	.34	.06
56	.36	.38	.30	.08
57	.52	.57	.35	.22
58	.59	.68	.34	.34
59	.40	.46	.24	.22
60	.41	.49	.19	.30

Table 3

Percentage Correct Scores for MRM

Item #	Total Group	Nonspeeded Group	Speeded Group	Difference
1	.77	.77	.78	-.01
4	.65	.65	.64	.01
7	.67	.67	.67	.00
10	.65	.65	.64	.01
13	.62	.62	.63	-.01
16	.83	.82	.84	-.02
19	.58	.58	.58	.00
22	.60	.60	.60	.00
25	.67	.67	.66	.01
28	.70	.70	.70	.00
31	.58	.58	.57	.01
32	.80	.80	.80	.00
33	.62	.61	.63	-.02
34	.62	.61	.62	-.02
35	.73	.72	.75	-.03
36	.63	.63	.64	-.01
37	.44	.44	.44	.00
38	.71	.71	.70	.01
39	.38	.37	.38	-.01
40	.51	.51	.50	.01
41	.65	.65	.65	.00
42	.60	.60	.61	-.01
43	.62	.62	.61	.01
44	.68	.68	.68	.00
45	.73	.72	.74	-.02
46	.54	.54	.53	.01
47	.59	.59	.59	.00
48	.56	.56	.57	-.01
49	.68	.69	.68	.01
50	.71	.71	.72	-.01
51	.59	.58	.60	-.02
52	.64	.64	.64	.00
53	.52	.52	.53	-.01
54	.63	.77	.20	.57
55	.36	.44	.13	.31
56	.33	.41	.08	.33
57	.55	.70	.11	.59
58	.70	.84	.29	.55
59	.44	.54	.12	.42
60	.46	.56	.17	.39

Table 4

Percentage Correct Scores for SGM(20, 2)

Item #	Total Group	Nonspeeded Group	Speeded Group	Difference
1	.71	.71	.73	-.02
4	.62	.62	.62	.00
7	.64	.64	.64	.00
10	.64	.64	.63	.01
13	.62	.63	.61	.02
16	.77	.77	.77	.00
19	.57	.57	.59	-.02
22	.59	.58	.60	-.02
25	.64	.64	.66	-.02
28	.67	.67	.68	-.01
31	.59	.58	.60	-.02
32	.76	.76	.77	-.01
33	.61	.61	.59	.02
34	.61	.61	.63	-.02
35	.68	.67	.70	-.03
36	.62	.61	.62	-.01
37	.48	.48	.50	-.02
38	.68	.68	.69	-.01
39	.43	.42	.46	-.04
40	.53	.53	.55	-.02
41	.64	.64	.63	.01
42	.59	.59	.60	-.01
43	.61	.61	.62	-.01
44	.66	.66	.68	-.02
45	.68	.68	.69	-.02
46	.56	.56	.57	-.01
47	.60	.60	.60	.00
48	.57	.57	.57	.00
49	.65	.66	.64	.02
50	.67	.68	.64	.04
51	.58	.59	.56	.03
52	.61	.62	.57	.05
53	.52	.54	.48	.06
54	.60	.63	.51	.12
55	.39	.40	.36	.04
56	.37	.38	.31	.07
57	.52	.57	.37	.20
58	.59	.67	.36	.31
59	.40	.45	.26	.19
60	.44	.49	.26	.23

Table 5

Percentage Correct Scores for SGM(9, 2)

Item #	Total Group	Nonspeeded Group	Speeded Group	Difference
1	.72	.71	.73	-.02
4	.62	.62	.63	-.01
7	.65	.64	.66	-.02
10	.63	.63	.63	.00
13	.61	.61	.62	-.01
16	.76	.76	.76	.00
19	.58	.58	.61	-.03
22	.59	.59	.60	-.01
25	.65	.65	.65	.00
28	.67	.67	.69	-.02
31	.58	.57	.60	-.03
32	.76	.76	.77	-.01
33	.60	.60	.60	.00
34	.62	.61	.62	-.01
35	.68	.68	.69	-.01
36	.61	.61	.60	.01
37	.48	.48	.48	.00
38	.66	.67	.65	.02
39	.42	.42	.43	-.01
40	.52	.52	.52	.00
41	.62	.63	.59	.04
42	.58	.58	.55	.03
43	.60	.61	.55	.06
44	.62	.64	.57	.07
45	.66	.69	.60	.09
46	.53	.55	.48	.07
47	.57	.59	.48	.11
48	.54	.56	.47	.09
49	.62	.65	.51	.14
50	.62	.67	.48	.19
51	.55	.59	.41	.18
52	.57	.63	.41	.22
53	.50	.54	.35	.19
54	.56	.63	.36	.27
55	.37	.40	.26	.14
56	.34	.37	.26	.11
57	.51	.58	.28	.30
58	.57	.67	.26	.41
59	.40	.46	.23	.23
60	.43	.50	.22	.28

Table 6

Location of First Speeded Item for Simulated Speeded Examinees

	Hybrid Model	MRM	SGM (20,2)	SGM (9,2)
Items 1-30	0	0	0	.011
31	0	0	0	.002
32	0	0	0	.008
33	0	0	0	.006
34	0	0	0	.004
35	0	0	0	.010
36	0	0	0	.009
37	0	0	.001	.012
38	0	0	.000	.018
39	0	0	.002	.013
40	0	0	.000	.020
41	0	0	.003	.028
42	0	0	.004	.023
43	0	0	.003	.029
44	0	0	.007	.036
45	0	0	.007	.027
46	0	0	.008	.037
47	0	0	.017	.055
48	0	0	.022	.046
49	0	0	.023	.050
50	0	0	.044	.069
51	0	0	.043	.049
52	0	0	.047	.054
53	0	0	.085	.070
54	.143	1.000	.088	.063
55	.143	0	.080	.052
56	.143	0	.158	.075
57	.143	0	.122	.044
58	.143	0	.106	.042
59	.143	0	.114	.034
60	.143	0	.016	.006

Table 7

Decrease in Percentage Correct Scores on First Speeded Item

Hybrid Model

First Speeded Item	E(% correct) on first speeded item[1]	% correct on first speeded item	Difference
54	.65	.23	.42
55	.44	.20	.24
56	.38	.24	.14
57	.59	.20	.39
58	.67	.20	.47
59	.51	.21	.30
60	.50	.18	.32

MRM Model

First Speeded Item	E(% correct) on first speeded item[2]	% correct on first speeded item	Difference
54	.77	.20	.57

[1]Expected percentage correct score for the hybrid model was estimated as the sum of 3PL probabilities of correct response across all examinees becoming speeded on item *i*.. [2]Expected percentage correct score for the MRM was estimated as the sum of $P(2|\$_{54,NS})$ across all examinees in the speeded group, where $\$_{54,NS}$ is the Rasch item difficulty for item 54 for the nonspeeded group.

Table 8

SGM(20,2)

First Speeded Item[1]	E(% correct) on first speeded item[2]	% correct on first speeded item	Difference
41	.63	.43	.20
42	.55	.67	-.12
43	.57	.57	.00
44	.65	.56	.09
45	.73	.50	.23
46	.56	.57	-.01
47	.58	.57	.01
48	.57	.44	.13
49	.66	.49	.17
50	.68	.53	.15
51	.59	.50	.09
52	.65	.53	.12
53	.55	.48	.07
54	.65	.52	.13
55	.45	.31	.14
56	.40	.37	.03
57	.61	.46	.15
58	.69	.48	.21
59	.47	.39	.08
60	.49	.48	.01

[1]Only items that were the first speeded item for at least 5 examinees are included. [2]Expected percentage correct score was estimated as the sum of 3PL probabilities of correct response across all examinees becoming speeded on item *i*.

Table 9

SGM(9,2)

First Speeded Item[1]	E(% correct) on first speeded item[2]	% correct on first speeded item	Difference
29	.66	.38	.28
30	.74	.17	.57
31	.64	.67	-.03
32	.77	.79	-.02
33	.61	.56	.05
34	.67	.60	.07
35	.69	.72	-.03
36	.64	.59	.05
37	.50	.45	.05
38	.69	.51	.18
39	.40	.33	.07
40	.56	.49	.07
41	.64	.43	.21
42	.59	.47	.12
43	.63	.50	.13
44	.65	.47	.18
45	.70	.66	.04
46	.56	.46	.10
47	.62	.47	.15
48	.58	.51	.07
49	.66	.58	.08
50	.70	.61	.09
51	.63	.48	.15
52	.64	.51	.13
53	.55	.45	.10
54	.63	.54	.09
55	.43	.36	.07
56	.39	.34	.05
57	.59	.55	.04
58	.68	.51	.17
59	.45	.38	.07
60	.56	.53	.03

[1]Only items that were the first speeded item for at least 5 examinees are included. [2]Expected percentage correct score was estimated as the sum of 3PL probabilities of correct response across all examinees becoming speeded on item *i*.

Table 10

$P_i^*(\theta_j)$ on End-of-Test Items for Simulated Speeded Examinees

		Item Number							
Model	θ	53	54	55	56	57	58	59	60
No Speededness	-2	.33	.30	.21	.20	.28	.38	.23	.32
	-1	.42	.43	.24	.22	.39	.52	.29	.39
	0	.53	.62	.35	.31	.56	.68	.43	.49
	1	.66	.80	.61	.60	.75	.82	.64	.60
	2	.78	.91	.86	.89	.89	.91	.83	.71
Hybrid	-2	.33	.29	.21	.20	.24	.25	.20	.20
	-1	.42	.40	.23	.21	.28	.29	.21	.20
	0	.53	.56	.31	.26	.36	.34	.23	.20
	1	.66	.71	.49	.43	.44	.38	.26	.20
	2	.78	.81	.67	.60	.50	.40	.29	.20
MRM	-2	.12	.03	.02	.01	.01	.04	.01	.02
	-1	.27	.07	.04	.02	.03	.11	.03	.06
	0	.50	.16	.10	.06	.08	.25	.08	.14
	1	.73	.34	.23	.15	.18	.48	.19	.30
	2	.88	.59	.45	.33	.37	.71	.39	.54
SGM (θ = .9)[1]	-2	.33	.30	.21	.20	.27	.34	.22	.28
	-1	.42	.43	.24	.22	.35	.44	.26	.33
	0	.53	.62	.34	.30	.50	.56	.36	.39
	1	.66	.80	.58	.55	.65	.67	.51	.46
	2	.78	.91	.82	.81	.76	.74	.65	.54
SGM (θ = .8)[1]	-2	.29	.27	.21	.20	.24	.29	.21	.25
	-1	.35	.35	.22	.21	.30	.35	.24	.28
	0	.44	.48	.29	.26	.39	.43	.30	.32
	1	.53	.60	.45	.43	.49	.50	.40	.37
	2	.61	.67	.61	.59	.56	.55	.59	.41

[1] $P_i^*(\theta_j)$ values for SGM were computed using δ = 3.912

Table 11

Change in $P_i^*(2_j)$ as a Function of $2, 0, 8$ and Item Number

		Item Number																			
		51		52		53		54		55		56		57		58		59		60	
	0	.9	.8	.9	.8	.9	.8	.9	.8	.9	.8	.9	.8	.9	.8	.9	.8	.9	.8	.9	.8
2	8																				
- 2	2.0	.30	.29	.30	.28	.30	.28	.30	.28	.29	.27	.29	.27	.29	.27	.28	.27	.28	.26	.28	.26
	4.0	.30	.28	.30	.27	.30	.27	.30	.26	.29	.26	.28	.25	.28	.25	.27	.25	.27	.24	.26	.24
	8.0	.30	.26	.30	.25	.30	.25	.30	.24	.28	.24	.27	.23	.26	.23	.25	.22	.25	.22	.24	.22
- 1	2.0	.42	.39	.42	.39	.42	.38	.42	.37	.41	.37	.40	.36	.39	.36	.39	.35	.38	.34	.37	.34
	4.0	.42	.38	.42	.36	.42	.35	.42	.34	.40	.33	.39	.32	.38	.31	.36	.30	.35	.30	.34	.29
	8.0	.42	.34	.42	.32	.42	.31	.42	.29	.39	.28	.36	.27	.34	.26	.32	.25	.31	.24	.29	.24
0	2.0	.60	.56	.60	.55	.60	.54	.60	.52	.59	.51	.57	.50	.56	.49	.55	.48	.54	.47	.52	.46
	4.0	.60	.53	.60	.50	.60	.48	.60	.46	.57	.44	.55	.43	.53	.41	.50	.39	.48	.38	.46	.36
	8.0	.60	.47	.60	.43	.60	.40	.60	.37	.55	.35	.50	.33	.47	.31	.43	.29	.40	.28	.37	.27
1	2.0	.78	.73	.78	.71	.78	.69	.78	.67	.77	.66	.75	.64	.73	.62	.71	.61	.69	.59	.67	.57
	4.0	.78	.68	.78	.64	.78	.61	.78	.58	.75	.56	.71	.53	.68	.50	.64	.48	.61	.46	.58	.44
	8.0	.78	.59	.78	.54	.78	.49	.78	.45	.71	.42	.65	.39	.59	.36	.54	.34	.49	.32	.45	.30
2	2.0	.90	.84	.90	.81	.90	.79	.90	.77	.88	.75	.86	.73	.84	.71	.81	.69	.79	.67	.77	.65
	4.0	.90	.77	.90	.73	.90	.70	.90	.66	.86	.63	.82	.60	.77	.57	.73	.54	.70	.51	.66	.49
	8.0	.90	.67	.90	.61	.90	.55	.90	.50	.82	.46	.74	.42	.67	.39	.61	.36	.55	.34	.50	.32

Note: $P_i^*(2_j)$ was computed using $\alpha = 1.0$, $\beta = 0.0$, and $c = 0.2$ for all items.