A comparison of Test Scoring Methods in the Presence of Test Speededness

Youngsuk Suh
Taehoon Kang
James A. Wollack
Su-Young Kim

Dept of Educational Psychology
University of Wisconsin-Madison
1025 W. Johnson, Room 859
Madison, WI 53706 USA

Paper presented at the Annual Meeting of the National Council on Measurement in Education (NCME), San Francisco, CA, April, 2006

A comparison of test scoring methods in the presence of test speededness

Introduction

There are many reasons why examinees might not provide answers to all items on a test. For example, answers may be omitted because an examinee carelessly and inadvertently skipped an item or entered the answer on the wrong line of the answer sheet. In such instances, the data are said to be missing at random, meaning that the probability that an observation (Y_{ij}) is missing is unrelated to the other missing observations after controlling all the observed data (Little & Rubin, 1987). Rubin (1976) proved that, under direct maximum likelihood and Bayesian estimation, data missing at random can be ignored, without affecting parameter estimation.

Other sources of nonresponse may be more systematic, such as when an examinee has insufficient time to consider answering the items. The effects of time limits on test performances have been referred to as *speededness effects* (Evans & Reilly, 1972). When examinees become speeded, their performance on items at the end of the test may change. Examinees who are hurried may find end-of-test items to be harder than do examinees with ample time, or may altogether fail to complete them. As a result, speededness has been shown to cause local item dependence (LID; Yen, 1993) among the items at the end of the test. Because of the systematic pattern of omitted responses among end-of-test items in speeded tests, these omitted responses cannot be classified as missing at random.

In the presence of test speededness, the parameter estimates of item response models can be poorly estimated, particularly for items located near the end of tests (Douglas, Kim,

Habing, & Gao, 1998; Oshima, 1994) and for slow-working examinees. Augmenting this perspective, Bolt, Cohen, and Wollack (2002) proposed a mixture Rasch Model (MRM) for reducing contamination in item difficulty estimates due to speededness, and concluded that the parameter estimates obtained for end-of-test items in the nonspeeded class more accurately approximated their difficulties when administered under non-speeded conditions. The MRM was extended to the 2-parameter logistic model (2PLM) and the 3-parameter logistic model (3PLM) by Bolt, Mroch, and Kim (2003).

In the past several years, a few models have been developed that model test speededness, in an attempt to improve the estimation of parameters for items at the end of the test (e.g., Yamamoto & Everson, 1997; Bolt et al., 2002; Bolt et al., 2003; Wollack, Cohen, & Wells, 2003). However, few simulation studies have been conducted to focus on the estimation of ability parameters when there are missing or abnormal responses due to test speededness. In spite of their success in accounting for test speededness, mixture item response theory (IRT) models have not previously been studied to evaluate ability estimation procedures for speeded tests with missing data. However, at least from a conceptual perspective, missing responses to end-of-test items would seem to be a rich source of information about the appropriateness of the time limits for certain individuals.

In this study, we examine the recovery of examinee ability parameters in the presence of speeded test data with some omitted responses among end-of-test items. Recovery is studied under three estimation procedures and two ways of scoring omitted responses. Data were simulated for three different distributions of examinee ability, two amounts of speededness, and two rates of omitting end-of-test items.

Methods

Simulation Design

The speededness-generating model (SGM; Wollack & Cohen, 2004) was used to simulate the responses affected by test speededness. This model is given by:

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

where $P_i(\theta_j)$ is the probability of examinee *j* answering item *i* correctly under the 2PLM, c_i is a pseudo-guessing parameter, and, $\eta_j (0 \le \eta_j \le 1)$ and $\lambda_j (0 \le \lambda_j)$ are the speededness point parameter and speededness rate parameter of examinee *j*, respectively. The function min(x, y) selects the smaller of the two values, x and y. η_j equals the percentage of items, proceeding from the beginning of the test, that will be completed before examinee *j* first experiences speededness. Therefore, smaller values of η_j correspond to examinee *j* becoming speeded earlier in the test. For example, $\eta_j = .75$ indicates that examinee *j* becomes speeded three-quarters of the way through the test. Once an examinee passes the speededness point, $\left[1 - \left(\frac{i}{n} - \eta_j \right) \right]$ is raised to the power λ_j, which serves to control the speed at which P_{ij}^{*} decreases. For (η_j, λ_j) pairs causing min(x, y) ≈ 0, P_{ij}^{*} approaches c_i. Examinees with $\eta_j = 1$ or $\lambda_j = 0$ are not speeded for any items. In such cases, P_{ij}^{*} reduces to the 3PLM.

As shown in Table 1, the design of the simulation study includes 12 conditions, resulting from three fully crossed factors: the location of the ability distribution, the amount of speededness (i.e., the percentage of examinees for whom the test is speeded), and the rate of

missing responses. The number of items and examinees in each condition were fixed as 100 and 2,000, respectively.

Insert Table 1 About Here

Item parameters (item discrimination (a_i), item difficulty (b_i), and item guessing (c_i)) and examinee parameters (η_j and λ_j) were generated from the following distributions for the 12 conditions.

$$a \sim \log normal\ (0,\ 0.5)$$

$$b \sim normal\ (0,\ 1)$$

$$c \sim beta\ (5,\ 17)$$

$$\eta \sim beta\ (120,\ 80)$$

$$\lambda \sim \log normal\ (3.912,\ 1)$$

Ability parameters were generated from normal distributions with a standard deviation of 1. However, the location of the mean of the distribution was manipulated to simulate groups of higher or lower ability. The mean values were set at either −1, 0, or 1. Also, it should be noted that by using $\eta \sim beta\ (120,\ 80)$, examinees were, on average, simulated to become speeded around the 61st item. The first simulated speeded item was between item 53 and item 67 for roughly 95% of the speeded examinees.

In addition, two amounts of speededness were simulated: low and high. The low speededness was simulated by having 30% of the total number of examinees as a speeded group, while the high speededness was simulated by having 70% of the examinees speeded. Among the

2,000 examinees, item responses for 600 (or 1,400) speeded examinees were generated using the above-specified parameters, while item responses for the remaining nonspeeded examinees were generated by fixing $\eta_j = 1$ and $\lambda_j = 0$, which indicates the 3PLM.

Finally, two different rates of missing responses due to test speededness were simulated: 20% and 40%. It should be noted that these values indicate the percentage of omitted responses among the items showing speededness, not among the total number of items. That is, for an examinee who is speeded on the last 40 items, responses for 8 of those items will be simulated as missing in the 20 % condition, and 16 will be missing in the 40 % condition. In terms of the total responses for all examinees, the rates of missing range from approximately 2.4% (40% speeded items x 20% omitted responses x 30% speeded examinees) to approximately 11.2% (40% speeded items x 40% omitted responses x 70% speeded examinees) of all responses. Missing responses were simulated to be among the last items on the test. However, because item parameters in a mixture IRT model cannot be estimated when all examinees in a particular group omit an item, missing responses were simulated using an odd-even approach. In this approach, each simulated examinee was issued an ID number in sequential order, beginning at 1. Responses for the last (20% or 40% of the) odd numbered items on the test were deleted for odd numbered speeded examinees, and responses for the last even numbered items were deleted for even numbered speeded examinees.

Estimation

Markov chain Monte Carlo (MCMC) algorithms have received increasing attention in IRT because they offer great promise in estimating more complex types of item response models (Baker, 1998; Patz & Junker, 1999a, 1999b; Kim, 2001; Wollack, Bolt, Cohen, & Lee, 2002)

and have been found to be particularly useful in deconvolving mixture distributions (Robert, 1996), including mixture item response models (Rost, 1990; Bolt, Cohen, & Wollack, 2001).

Each simulated dataset was analyzed by three different estimation algorithms. Estimation of the 3PLM parameters was done both using marginal maximum likelihood (MML; Bock & Lieberman, 1970) estimation and Bayesian estimation. MML estimates of item parameters, hereafter referred to as the 3PL-MML condition, were obtained using the computer program MULTILOG (Thissen, Chen, & Bock, 2003). Bayesian estimates were obtained using MCMC algorithms, hereafter referred to as the 3PL-MCMC condition, implemented with the computer program WinBUGS 1.4 (Spigelhalter, Thomas, Best, & Lunn, 2003). Finally, a mixture 3PLM (M3PLM) was used, based on a MCMC algorithm using WinBUGS. To estimate parameters under the 3PL-MCMC and M3PLM conditions, each model parameter was sampled from its full conditional distribution 10,000 times. The initial 5,000 draws were discarded for the burn-in. Parameter estimates were taken as the mean of the remaining sampled values. Group membership was estimated for each examinee as the modal group (speeded or nonspeeded) among all sampled values after burn-in.

Each of the three estimation algorithms was applied twice to each dataset, once by scoring missing data as incorrect and once by treating missing data as omitted. Therefore, six different procedures of estimating examinee ability were used for each condition.

Evaluative measures

Several different evaluative measures were obtained to investigate the relative performance of ability estimation procedures. Root mean square errors (RMSE) and Pearson correlations were computed between generating and estimated ability parameters for all six

estimation procedures. The correlations were obtained both for the whole examinee group and for the two separate groups (nonspeeded and speeded).

Results

The means and standard deviations of the raw scores for 12 data sets are shown in Table 2. Mean scores increased as the mean of the generating ability distribution increased. Also, mean scores were higher when 30% of the examinees were speeded than when 70% of the examinees were speeded. The standard deviations ranged from 10.91 to 15.87. Standard deviations also tended to increase with the mean of the ability distribution, but was relatively unaffected by the amount of speededness. The Cronbach α ranged from .83 to .93, and increased as mean ability increased.

Insert Table 2 About Here

The proportion of correct group membership classification for the two scoring methods under the M3PLM is shown in Table 3. The range varied from 84% to 99.8%. The average for the omitted scoring was 93.9%, while the average for the incorrect scoring was 96.2%. Therefore, in general, scoring missings as incorrect recovered the underlying group memberships slightly better than scoring missings as omitted.

Insert Table 3

RMSE was calculated as the square root of the average squared difference between the estimated and true ability parameter, after first scaling them onto a common ability metric. Because IRT parameters are invariant up to a linear transformation, solutions from different models and across different datasets will not be comparable unless they are first transformed to a common scale. Because the purpose here was to assess parameter recovery, the scale of the underlying generating parameters was taken to be the base scale to which all others were equated. Parameters were equated to the base scale using the test characteristic curve method (Stocking & Lord, 1983) as implemented in the computer program EQUATE (Baker, Al-Karni, & Al-Dosary, 1991). To ensure that only nonspeeded items were used for equating (Wollack et al., 2003), items 1 through 50 were used as the anchor items for purposes of estimating the equating coefficients. These coefficients were then used to transform all ability estimates onto the base scale.

Table 4 illustrates the RMSE values for the different estimation procedures, which ranged from 0.468 to 1.486. For the 3PL-MML and 3PL-MCMC models, the omitted scoring consistently showed smaller RMSEs than the incorrect scoring. However, for the M3PLM, scoring missings as omits produced smaller RMSEs when the mean of the ability distribution was −1, but larger RMSEs when the mean was 0.0 or 1.0. Across the 12 conditions, scoring missing data as incorrect under the M3PLM showed the smallest RMSE, averaging just 0.567. The M3PLM's RMSE for the omitted scoring was somewhat larger, 0.621, but was still considerably smaller than the RMSEs for either of the other two models. On the other hand, the RMSE for the 3PL-MML scoring missing as incorrect was 0.976, the largest among the six scoring methods.

The mean of the ability distribution appeared to influence the quality of recovery. RMSE values were smallest when ability was sampled from a distribution with mean of 0.0. This is not surprising because the item difficulties were also generated with a mean of 0.0. Also, in general, RMSEs were larger in the N(1, 1) condition than in the N(−1, 1) condition. This may be attributable to the distribution of item difficulty parameters in the simulated 100-item test. Though the test was of average difficulty (mean = 0.01), five of the seven most extreme items, and the three most extreme items, were all negative.

Insert Table 4 About Here

For the 3PLM in both estimation algorithms (MML and MCMC), the differences between omitted and incorrect scoring methods tended to be large. In particular, this tendency was severe when there were a smaller number of examinees (30% amount of speededness) in the speeded group and a large number of speeded items were left missing (40% rate of missing). The difference between the omitted and incorrect scoring methods appeared less severe in the presence of 70% amount of speededness than a 30% amount of speededness.

Tables 5 and 6 show the Pearson correlations between the estimated and true ability parameters. Table 5 shows the correlations for the whole examinee group, while Table 6 shows the correlations separately for the nonspeeded and speeded groups. In Table 5, the correlations range from 0.390 to 0.944. Similar patterns in the results of RMSE were found. In general, the omitted scoring consistently showed larger correlations than the incorrect scoring regardless of which estimation method was used. There were no substantial differences among the three ability distributions.

The correlations from the omitted scoring under the M3PLM appeared to be the largest among the six scoring methods across all 12 conditions, showing the range from 0.896 to 0.944. For the M3PLM, correlations looked very similar for the two scoring methods (omitted and incorrect) across all 12 conditions, though the omitted scoring conditions provided higher correlations in most cases. The 3PLM correlations (both MML and MCMC) were poor when missing responses were scored as incorrect, but were smallest under the 3PL-MML. Also, for the 3PLM, the differences between omitted and incorrect scoring methods tended to be large when 40% of items were missed among the speeded items. In particular, this tendency was severe when there were a small number of examinees in the speeded group (30% amount of speededness). By comparing two amounts of speededness for the 3PL-MML and 3PL-MCMC, the differences between omitted and incorrect scoring procedures were less severe in the 70% amount of speededness condition than in the 30% condition.

Insert Table 5 and 6 About Here

As shown in Table 6, the within-group correlations varied from 0.150 to 0.957. It should be noted that the correlations for all three models were based on the true group membership. The correlations for the M3PLM based on estimated group membership were slightly lower than those based on the true membership (the average difference was just 0.004). Consistent with the results in Tables 5, the omitted scoring under the M3PLM showed larger correlations than the other scoring procedures, while the incorrect scoring under the 3PL-MML showed the smallest correlations, on average. The differences between the omitted and incorrect scoring under the M3PLM appeared to be very small for all 24 cases.

In general, the nonspeeded groups showed similar correlations across the six scoring methods. This was especially true when there were large number of examinees in nonspeeded group (30% amount of speededness) and there was a small number of missing responses (20% missing rate), the correlations for the nonspeeded groups were nearly identical to each other for the six scoring procedures. Also, when there were a large number of examinees in speeded groups (70% amount of speededness), the correlations for the speeded groups were similar across the scoring procedures. The severe differences were detected for speeded groups in the 30% amount of speededness conditions. In particular, when a large number of missing responses (40% rate of missing) were present, the differences among the six scoring methods were substantial. The lowest correlations were for the speeded group under the 3PL-MML when the amount of speededness was 30 %.

Conclusion and Discussion

Results from this simulation study would appear to have several implications for how practitioners choose scoring procedures for missing responses in the presence of test speededness. First, the M3PLM recovered the underlying ability metric better than either of the other two estimation techniques. Second, the omitted scoring generally worked better than the incorrect scoring. By combining the two results, the omitted scoring under the M3PLM appeared generally to be the best though there were some conditions for which scoring missing as incorrect led to better estimation in the M3PLM. Uniformly, the incorrect scoring under the 3PL-MML appeared to be the worst method. Third, for the 3PLM, the differences between omitted and incorrect scoring procedures tended to be large when a large number of items were missing among the speeded items. In particular, when there were a small number of examinees

in the speeded group, the difference appeared to be more severe. Fourth, when there is a large amount of speededness (70% amount of speededness), there is relatively small difference between omitted and incorrect scoring in terms of ability estimation. Fifth, there were small differences among the three ability distributions, with recovery being slightly improved when the ability distribution matched the mean of the item difficulty distribution. Seventh, the six scoring procedures showed similar results in terms of ability metric recovery for nonspeeded groups. On the other hand, the two 3PL algorithms did not work properly for speeded groups when missing responses were treated as incorrect. For estimating ability parameters of speeded group, the scoring procedure should be selected with caution. Finally, when scoring missings as incorrect, this study replicates the findings of Bolt et al. (2003) in that there were very large differences between the M3PLM and the 3PLM models. However, it is interesting to note that the differences between the M3PLM and the 3PLM were much less pronounced when missings were treated as omitted. Still, it appears as though the differences are significant enough to warrant using a M3PLM if speededness effects are believed to exist.

When a test is suspected to be speeded to some degree, important consideration must be given to the scoring of omitted or otherwise speeded items. Even though the evaluation of ability estimation procedures is very important for the successful application of IRT, results from different treatments of missing data in estimating ability, particularly systematically missing data such as for speeded tests, have not been fully studied. Recommendations from this study should help test developers and measurement specialists select and better understand test scoring procedures for working with incomplete data due to test speededness.

Limitations

Fixed numbers of examinees and items were used in this simulation study. The impact of varying these, particularly by choosing different sample sizes, needs to be considered in future studies. SGM (Wollack & Cohen, 2004) was used to simulate the responses affected by test speededness. However, reasonable and realistic distributions for η_j and λ_j are still being established. Similarly, other models (e.g., hybrid model of Yamamoto & Everson (1997)) could have been used to generate speeded datasets, and more realistic approaches for simulating systematically missing responses need to be investigated. Finally, this study represents a fairly extreme scenario which is unlikely to occur in practice. It would be fairly uncommon to encounter an educational test that was speeded for as many as 70% of the examinees and that became speeded only 60% of the way into the test. However, by studying the robustness of models under such extreme situations, it is possible to gain an appreciation for their differences. The application of these models to real data and to more subtle amounts of speededness will be an area for future exploration.

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.

Baker, F. B. (1998). An Investigation of the Item Parameter Recovery Characteristics of a Gibbs Sampling Procedure. *Applied Psychological Measurement, 22,* 153-169.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika, 35,* 179-197.

Bolt, D.M., Cohen, A.S., & Wollack, J.A. (2001). A mixture model for multiple choice data. *Journal of Educational and Behavioral Statistics, 26,* 381-409.

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. (2003, April). *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.

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

Evans, F. R., & Reilly, R. R. (1972). A study of speededness as a source of test bias. *Journal of Educational Measurement, 9,* 123-131.

Kim, S.-H. (2001). An evaluation of a Markov chain Monte Carlo method for the Rasch model. *Applied Psychological Measurement, 25,* 163-176.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical Analysis with Missing Data.* John Wiley and Sons.

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.

Robert, C. P. (1996). Mixtures of distributions: Inference and estimation. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 441-464). Boca Raton, FL: Chapman & Hall.

Rost, J. (1990). Rasch model in latent cases: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14,* 271-282.

Rubin , D. B. (1976). Inference and missing data. *Biometrika, 63,* 581-592.

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Lunn, D. (2003). WINBUGS 1.4* User Manual. [Computer Program.]

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201-210.

Thissen, D., Chen, W-H, & Bock, R.D. (2003). *Multilog (version 7)* [Computer software]. Lincolnwood, IL: Scientific Software International.

Wollack, J. A., Bolt, D. M., Cohen, A. S., & Lee, Y. -S. (2002). Recovery of item parameters in the nominal response model: A comparison of marginal maximum likelihood estimation and Markov chain Monte Carlo estimation. *Applied Psychological Measurement, 26,* 339-352.

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.

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

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. Simulation design

Ability Distribution	Amount of Speededness	Rate of Missing
$\theta \sim N(-1, 1)$	30% (600 examinees)	20%
		40%
	70% (1400 examinees)	20%
		40%
$\theta \sim N(0, 1)$	30% (600 examinees)	20%
		40%
	70% (1400 examinees)	20%
		40%
$\theta \sim N(1, 1)$	30% (600 examinees)	20%
		40%
	70% (1400 examinees)	20%
		40%

Table 2. Descriptive statistics based on raw scores

Ability Distribution	Amount of Speededness	Rate of Missing	Mean	Standard Deviation	Cronbach α
$\theta \sim N(-1, 1)$	30%	20%	44.71	13.02	.88
		40%	44.26	13.30	.88
	70%	20%	39.76	10.91	.83
		40%	38.69	11.13	.84
$\theta \sim N(0, 1)$	30%	20%	56.36	15.10	.91
		40%	55.88	15.44	.92
	70%	20%	49.30	13.03	.88
		40%	48.20	13.34	.89
$\theta \sim N(1, 1)$	30%	20%	67.61	15.43	.93
		40%	67.13	15.87	.93
	70%	20%	58.62	14.23	.91
		40%	57.53	14.66	.92

Table 3. Correct classification of group membership in M3PLM

Ability Distribution	Amount of Speededness	Rate of Missing	Scoring of Missing	
			Omitted	Incorrect
$\theta \sim N(-1,\ 1)$	30%	20%	88.6%	95.2%
		40%	86.8%	96.6%
	70%	20%	90.1%	90.5%
		40%	88.0%	84.0%
$\theta \sim N(0,\ 1)$	30%	20%	95.7%	98.3%
		40%	94.1%	98.6%
	70%	20%	95.8%	97.9%
		40%	95.1%	95.3%
$\theta \sim N(1,\ 1)$	30%	20%	98.8%	99.7%
		40%	98.2%	99.8%
	70%	20%	98.0%	99.1%
		40%	97.2%	99.4%

Table 4. Root Mean Square Errors

Ability Dist.	Amt. of Speed	Rate of Missing	3PL-MML		3PLM-MCMC		M3PLM-MCMC	
			Omitted	Incorrect	Omitted	Incorrect	Omitted	Incorrect
$N(-1, 1)$	30%	20%	0.766	0.848	0.654	0.797	0.523	0.621
		40%	0.749	1.427	0.714	1.049	0.528	0.625
	70%	20%	0.744	0.833	0.647	0.710	0.521	0.642
		40%	0.747	0.829	0.786	0.761	0.536	0.677
$N(0, 1)$	30%	20%	0.678	0.963	0.627	0.987	0.515	0.468
		40%	0.597	0.987	0.588	1.016	0.514	0.469
	70%	20%	0.606	0.781	0.679	0.703	0.545	0.488
		40%	0.674	0.851	0.664	0.749	0.533	0.489
$N(1, 1)$	30%	20%	1.121	1.134	0.926	1.486	0.776	0.569
		40%	0.739	1.133	0.961	1.275	0.795	0.565
	70%	20%	0.744	0.962	0.994	0.900	0.820	0.594
		40%	0.904	0.965	0.979	0.917	0.849	0.591
		mean	0.756	**0.976**	0.768	0.946	0.621	**0.567**

Table 5. Pearson correlations for the whole examinee group

Ability Dist.	Amt. of Speed	Rate of Missing	3PL-MML		3PLM-MCMC		M3PLM-MCMC	
			Omitted	Incorrect	Omitted	Incorrect	Omitted	Incorrect
$N(-1, 1)$	30%	20%	.861	.649	.863	.763	.910	.907
		40%	.878	.424	.879	.443	.910	.905
	70%	20%	.833	.764	.848	.804	.896	.888
		40%	.848	.690	.855	.745	.896	.871
$N(0, 1)$	30%	20%	.834	.462	.832	.438	.944	.943
		40%	.914	.423	.861	.390	.944	.943
	70%	20%	.914	.693	.794	.765	.929	.927
		40%	.798	.644	.805	.712	.929	.926
$N(1, 1)$	30%	20%	.488	.422	.640	.439	.938	.937
		40%	.908	.394	.704	.415	.939	.939
	70%	20%	.910	.597	.655	.649	.916	.916
		40%	.654	.568	.666	.621	.919	.918
		mean	.820	**.561**	.784	.599	**.923**	.918

Table 6. Pearson correlations for nonspeeded and speeded groups

Ability Dist.	Amt. of Speed	Rate of Missing	Group	3PL-MML		3PLM-MCMC		M3PLM-MCMC	
				Omitted	Incorrect	Omitted	Incorrect	Omitted	Incorrect
$N(-1, 1)$	30%	20%	NS	.925	.910	.926	.920	.924	.923
			S	.862	.644	.866	.866	.890	.890
		40%	NS	.926	.893	.927	.893	.924	.921
			S	.871	.155	.874	.225	.891	.888
	70%	20%	NS	.903	.822	.892	.861	.924	.928
			S	.864	.834	.869	.860	.886	.879
		40%	NS	.914	.772	.898	.815	.924	.921
			S	.871	.810	.872	.849	.885	.871
$N(0, 1)$	30%	20%	NS	.946	.943	.948	.941	.957	.957
			S	.865	.185	.869	.219	.919	.918
		40%	NS	.913	.942	.953	.940	.957	.957
			S	.916	.150	.885	.168	.920	.918
	70%	20%	NS	.913	.798	.875	.814	.954	.954
			S	.914	.841	.872	.870	.919	.916
		40%	NS	.882	.757	.886	.756	.954	.955
			S	.875	.828	.876	.862	.919	.916
$N(1, 1)$	30%	20%	NS	.944	.943	.931	.943	.951	.952
			S	.289	.211	.733	.455	.910	.906
		40%	NS	.907	.942	.932	.941	.951	.951
			S	.910	.172	.783	.471	.913	.911
	70%	20%	NS	.909	.725	.789	.678	.952	.953
			S	.911	.788	.810	.823	.900	.900
		40%	NS	.781	.722	.802	.632	.953	.954
			S	.810	.782	.816	.823	.904	.903