

Chapter 7

Modeling the Effects of Test Length and Test Time on Parameter Estimation Using the HYBRID Model

Kentaro Yamamoto and Howard Everson

The Educational Testing Service, Princeton
The College Board, New York

1. Introduction

The speed of performing a task is one of the more noticeable aspects in which individuals differ from each other, in addition to the differences in the abilities to perform a task correctly. Traditional methods of assessing test speededness are limited to analyses of distributions of missing responses, especially at the end of a test. However, analysis of missing responses is inadequate in evaluating the speededness of a multiple choice test when the test score is a function of the total number of correct responses. With such a test, the most sensible strategy for an examinee who is running out of time is to quickly guess the answers to the remaining questions and increase his total number of correct responses by chance alone. Disregarding this widespread strategy when analyzing data from timed examinations surely underestimates the actual speededness of the test, and will likely result in biased ability estimates. In the past, many ability measurement models such as IRT did not explicitly incorporate speededness into the construct of ability. Hence, the construct of ability and the parameters in the model are assumed to be unaffected by the variation of the time limit in which the test is administered. It is common, however, for the performance level to decline if not enough time is allocated to the task. In such a case, analysis of missing responses alone is not adequate to measure performance under speeded conditions. Therefore, in order to obtain a more accurate assessment of speededness, it is necessary to evaluate not only missing responses, but also random responses. In this paper we demonstrate how the HYBRID model (Yamamoto, 1990) can be used to assess the effects of test speededness on both examinee ability and item parameter estimates.

Two previous studies (Bejar, 1985; Secolsky, 1989) attempted to examine speededness beyond counting omitted responses. The shared shortcomings in these earlier studies are: 1) they did not study the performance of their procedures when the test was not speeded; 2) they examined the speededness of the test at an arbitrary point in the test; 3) they were unconcerned with the bias of the IRT model parameters due to the speededness of the test; and 4) they did not assess the presence of differential speededness by subgroups of examinees, whatever the „subgroup“ definitions may be.

Recent developments in item response modeling address these problems, namely, the extended HYBRID model (Yamamoto 1990, 1995) for speededness of the test. The HYBRID model changes the question concerning speededness from „Is a test speeded?“ to „How speeded is a test?“ The HYBIL computer program, that is available from the first author, estimates the proportion of examinees who switch to a guessing strategy at each item sequentially in the test. When an examinee switches to a guessing strategy midway through a test, the probabilities of making correct responses on the following items no longer adhere to

the IRT model. Such a change in conditional probability is most noticeable among more able examinees. The HYBRID model expands the notion of the speededness of a test to include changes in conditional probabilities, while also analysing omitted responses. In addition, this model-based approach examines the effects of speededness on the estimated item parameters. Indeed, this approach is particularly useful for tests in which item difficulties are gradually increased.

Background of the HYBRID model

Traditional IRT (including unidimensional and multidimensional) and classical test theories use a single model to describe the behavior of all examinees. The HYBRID model by Yamamoto (1987, 1989), see also chapter 2.5, describes the mixture of examinees whose responses can be characterized by either an IRT based (a class with variation of individual ability) or a latent class based (multinomial independent class) item response model. The HYBRID model uses multiple item response models to describe the behavior of all examinees, while one model per examinee is posited. The HYBRID model can be extended to the case of strategy switching; i.e., where a subset of responses of an examinee is best described by a latent class (a guessing class) while IRT is most appropriate for the rest of the responses.

The model offers three advantages: 1) it characterizes examinees' strategy use, when salient; 2) it detects extraneous strategy influences in estimated model parameters; and 3) it provides an opportunity to incorporate partial knowledge of latent classes. The extended HYBRID model attempts to provide of qualitative information of the knowledge possessed by examinees by relying more on the qualitative aspects of the examinees' cognitive characteristics-by-items interaction, i.e., the interaction of the test taking speed of examinees with the location of items in the test. This occurs often in speeded tests when examinees run out of time and switch from a thoughtful response strategy to a strategy of patterned or random responses. Standard IRT cannot handle this phenomenon and can often yield misleading inferences about the proficiencies of the examinees and the properties of the test items.

2. The Model

Assumptions of the extended HYBRID model are: 1) under a patterned response strategy, the conditional probability of a correct response is independent of one's ability; 2) the response to an item by each examinee can be characterized either by an IRT model of a particular form or a patterned response model; and 3) conditional independence holds, propensity of correct response is constant given item parameter, ability and strategy are the same.

In many large scale assessments and achievement tests, the great majority of omitted responses are found at the end of the tests. The HYBRID model is useful in such cases. However, this switch-only-once assumption is not very rigid. Indeed, by using a probabilistic model minor deviations from this modeling structure can be tolerated, especially when switching occurs early in item sequences. As noted earlier, the model is probabilistic and considers all possible switching points for every examinee. Hence, deviations from the model's assumption have a minimal effect on the model's overall probability structure. For example, if an examinee taking a prototypical reading comprehension test skips the next to last passage and goes on to the last (and perhaps most difficult) passage, the model would estimate nearly equal probability of switching anywhere between the penultimate passage

and the end of the test if the examinee had low ability. In other words, it would not provide a precise point of switching to random responses. This is largely because the conditional probability based on ability is nearly identical to the probability of randomly selecting correct responses for the low ability examinees. However, if the examinee had high ability, and skipped the next to last passage but correctly answered questions in the last passage, the model would detect two main probable switching locations, one at the beginning of the penultimate passage and the other at the end of the exam.

The following function expresses the likelihood of a response vector \mathbf{x}_v based on the propensity of correct response on an item i , and group g . The notation indicates that up to m_g th item the IRT model holds and a multinomial model would be appropriate for the rest of items. Thus the two distinct models are applicable to a single examinee, that is, two different models hold in distinct parts of the response pattern.

$$p(\mathbf{x}_v) = \sum_{g=1}^G \pi_g \prod_{i=1}^{m_g} \frac{\exp(x \beta_i (\theta_v - \sigma_i))}{1 + \exp(\beta_i (\theta_v - \sigma_i))} \prod_{i=m_g+1}^k \pi_{ixg}$$

The HYBRID model incorporates IRT parameters for items and examinees' abilities as well as parameters that define the distribution of the examinees who switched response strategies. The earlier version of the HYBRID model estimated IRT parameters and latent class parameters simultaneously, i.e., the distribution of subjects in the latent classes and the conditional probabilities of a correct response given a latent class. In the extended model, the proportion of subjects who switch strategies on an item is estimated along with the IRT parameters, while fixing the conditional probabilities of a correct response π_{ixg} for the patterned response. The marginal maximum likelihood method to estimate IRT parameters developed by Bock and Aitkin (1981) is used to estimate model parameters, and incorporated into a computer program HYBIL (Yamamoto, 1989). In many cases a non-informative prior distribution for the bivariate distribution of switching behavior and ability $f(\theta, g)$ may be used, except when $g = k$, and $f(\theta|g)$ has a standard normal distribution. It is feasible to incorporate more constrained distributional forms that may lead to more stable results under some conditions. One reasonable distribution may be that each $f(\theta|g)$ is normal with a mean that is a function of g and a common variance, and the marginal distribution of g divided by the number of items has a beta distribution. Supposing the mean of $f(\theta|g)$ being an increasing function of g enables us to represent the common observation of „more able respondents complete tests than less able respondents“. The model assumes that all tests are speeded to various degrees, in a range of hardly speeded to very speeded with extremes represented by beta distribution for g .

In an effort to demonstrate the efficacy of the HYBRID model for detecting test speededness by identifying strategy switching, we applied the model to two different sets of examinee data, one simulating data on 3,000 examinees taking a 70 item test, and the other on data gathered as part of a field study of 752 examinees taking a 40 item multiple choice reading comprehension test in which allotted testing time was varied. The extended HYBRID model was used to analyze both sets of data, and the simulated and actual strategy switch points were mapped onto the structure of the test.

3. Simulation study

The simulated data set consists of 3000 ability parameters from a standard normal distribution $N(0,1)$, and 70 pairs of item parameters of the standard 2PL IRT model simulated

from independent normal distributions, $N(1, 0.4)$ for β , and $N(0.0, 0.8)$ for σ . Based on these simulated parameters, a 3000×70 response matrix was generated. Number of simulees switching to random responses were increased in increments of 50 starting with the 51st item. Thus among 3000 simulees, 2000 did not switch to random responses ($m_g = 70$), while 50 simulees switched to random response at the 51st item ($m_g = 50$), 50 more at the 52nd item ($m_g = 51$), so on till the last (70th) item ($m_g = 69$). Responses of simulees switching to random response were generated based on $\pi_{ixg} = 0.2$.

<i>RMSD of item parameter estimates against option (3)</i>				
Parameters	(1) IRT only		(2) HYBRID	
	Mean Deviation	RMSD	Mean Deviation	RMSD
Items 1-70				
Slope β	.02	.22	.02	.03
Location σ	.09	.24	-.06	.08
Items 1-50				
Slope β	.02	.03	.02	.02
Location σ	.00	.03	-.06	.06
Items 51-60				
Slope β	.21	.25	.00	.01
Location σ	.22	.33	-.05	.06
Items 61-70				
Slope β	.48	.51	.01	.05
Location σ	.39	.53	-.11	.14
Ability (simulees)				
1-2000	.03	.06	-.00	.05
2001-2500	-.03	.08	-.01	.07
2501-3000	-.10	.21	-.00	.09
2501-3000, $\theta < 0$.02	.08	-.02	.06
2501-3000, $0 < \theta < 1$	-.02	.07	-.02	.05
2501-3000, $1 < \theta < 2$	-.34	.25	.04	.10
2501-3000, $2 < \theta$	-1.0	.61	.22	.19
Fit of the model				
-2*log-likelihood	189,731		185,348	
No. of Parameters	140		229	
AIC	190,011		185,806	

Table 1: Fit of the model and accuracy of estimated parameters by the HYBRID model and the 2PL IRT model

Three sets of model parameters were estimated on the simulated data: 1) ordinary 2PL IRT parameters (140 parameters); 2) HYBRID model parameters ($140 + 60 + 29 = 229$ parameters); and 3) ordinary 2PL IRT parameters with random responses treated as not presented (140 parameters) serving as the best possible estimates based on this data set.

The -2*log-likelihood for the IRT model on the omit data is not comparable due to the fact that 300×10 responses were never included in calculating the likelihood; hence, it is not reported here. The mean deviation is calculated as the mean of the differences between parameter estimates for option 1 and 2, respectively, and the estimate for option 3, and

$$\text{RMSD} = \sqrt{\sum \text{Dev}^2}.$$

For the HYBRID model parameter estimation, $f(\theta|g)$ was constrained to be normal and considered only for the last 30 items as it was opted instead of entire 70 items, making 60 parameters to be estimated. In addition, 29 π_g parameters were to be estimated. The rationale here is that when item parameters are estimated using this third option, the IRT item parameter estimation is dependent only on the portion of the data which correspond to the IRT model, hence the estimation error is due to randomness of the data. The IRT item parameters estimates based on the competing models would be compared against the estimates of this third option. Summary statistics of mean deviations and root mean square deviations comparing the estimated item and ability parameters, and the values of $-2 \times \log$ -likelihood to indicate the fits of the models are presented in Table 1.

Note that the random responses started after the 50th item, and estimation errors are negligible until the 50th item. Error increases for items located later in the item sequence, i.e., those with greater proportions of random responses. The estimated π_g 's showed the clear demarcation of the 51st item where random responding started. Slight over-estimation of π_g 's before 50th item (about 1%) and near the 70th item (about 3%) were found. The cumulative distribution of switched population was presented in Figure 1. It shows that the estimate closely follow the real distribution indicated by the solid line.

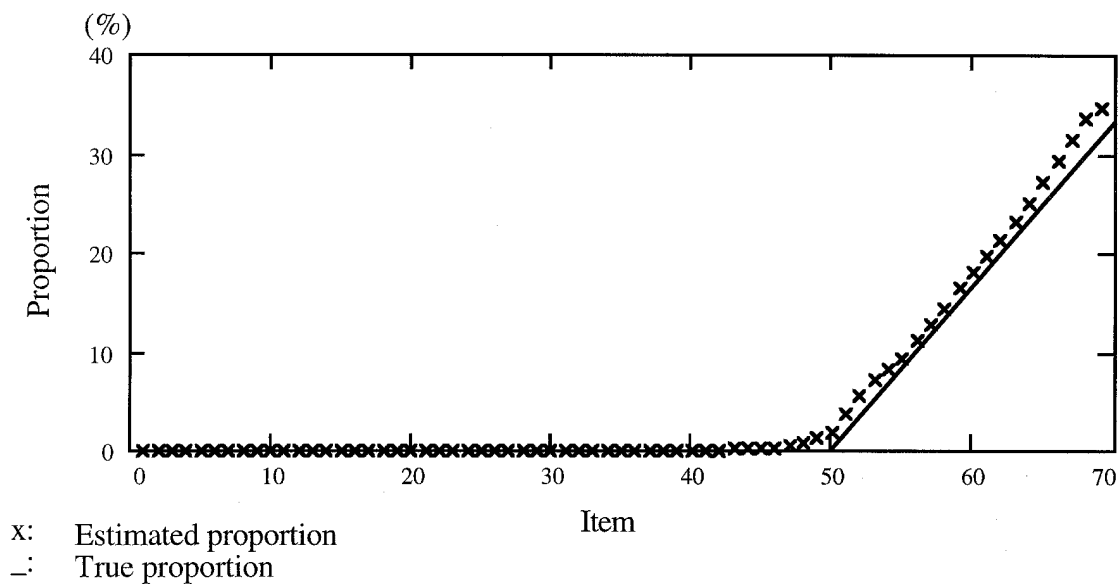
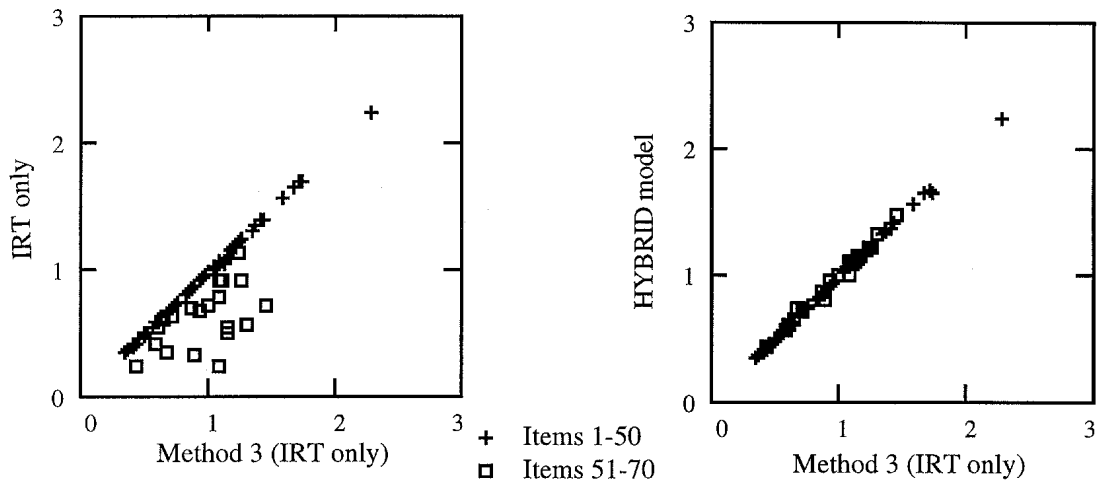


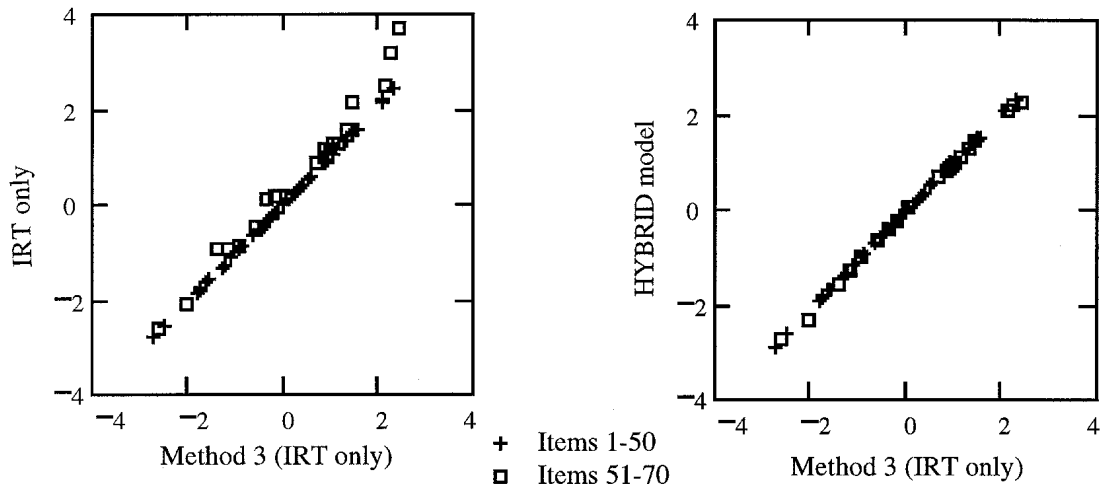
Figure 1: Cumulative proportion of speeded population

Two sets of estimated item parameters, one from the IRT only model, and the other from the HYBRID model are plotted against estimated parameters using the option 3 (Figure 2).

Slope Parameter



Location Parameter



Ability Parameter

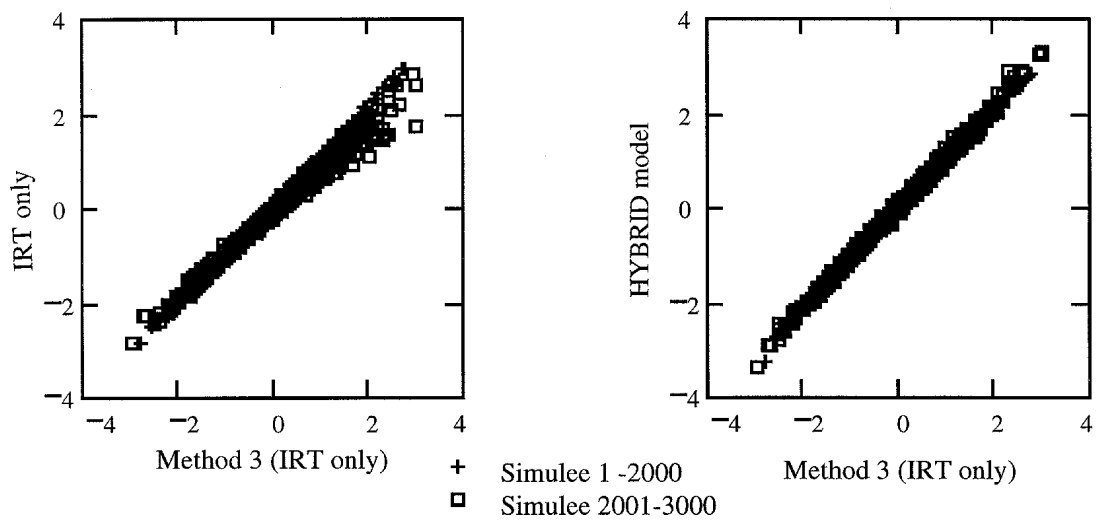


Figure 2: Comparison of estimated IRT parameters for the IRT only and the the HYBRID models

It is quite clear that HYBIL successfully eliminated the influence of the random response subpopulation on the estimated item parameters. The estimated item parameters of the last ten items are clearly set apart from the rest of the item parameters based on a comparison among the results from the IRT only estimation. The RMSDs for several sets of items, such as last 10 items which included 17% of random responses, are presented in Table 1, and they clearly indicate the inaccuracy of parameter estimates when random responses are ignored. However, it is significant to note that the conventional model fit statistics cannot detect the presence of random responses in the data set due to the speededness of the test. The overall accuracy of the estimated item parameters was markedly improved using the HYBRID model. Moreover, the accuracy of the ability estimates was improved using the HYBRID model, especially for those patterns that included at least 10 random responses. The under estimation of ability for very able simulees is more severe under the option 1 of the IRT only method, (e.g., 0.02 for those with $\theta < 0.0$, -0.02 for $0 < \theta < 1$, -.337 for $1 < \theta < 2$, and finally -1.0 for $\theta > 2$.)

The competing models were also evaluated in terms of their ability to identify lack of speededness of the test when the test was not speeded. This was done by using the original IRT only data before random responses replaced the part of the data. The two methods of (1) the IRT only and (2) the HYBRID model produced nearly identical results in estimated parameters and model fit, and the cumulative distribution of switched proportion only amounted to 0.4% at the last item. Thus the HYBRID model correctly identified that the data as not speeded in this simulated data set.

It is clear that ignoring random responses during IRT parameter calibration has serious impact on item parameters and ability parameters. The HYBRID model effectively eliminates such impact and produced corrected parameters.

4. Field Study

The field study was conducted to examine the utility of the extended HYBRID model for analyzing test data and detecting speededness for different subgroups of the test taking population - English as a second language (ESL) and English as a primary language (EPL) examinees - when tests are administered under different time conditions. Data for the field study came from an administration of a forty item multiple-choice reading comprehension test administered to 752 examinees. The forty items in the test were based on ten reading passages of varying lengths, and the scores are based on the total number correct. A quasi-experimental research design was used in which groups of examinees were administered the test under differently timed conditions. Students from a large urban university took the test in groups of thirty that were randomly assigned to either a 45 minute or a 60 minute test condition. The sample was ethnically diverse, consisting of 40% African Americans, 17% Hispanics, 12% Asian Americans, 12% Whites, and 19% unclassified. Moreover, 16% self-identified as students for whom English was a second language.

The extended HYBRID model was used in the analysis of these data in an attempt to map the switch points on to the structure of the test. Since this test contained a number of brief reading passages followed by a short series of multiple choice items, mapping the points where examinees affected by speeded “switch” to random response patterns could be achieved. This application of the HYBRID model permits us to test its utility for detecting the effects, if any, of the different time limits on both test speededness and the ability estimates for different subgroups of examinees, as well as classifying the structure of the test under the two timed conditions. Thus the analysis focused on the cumulative proportions of

EPL and ESL examinees who switched to a random response strategy in the 45 minute and 60 minute test conditions. Table 2 shows the cumulative proportions of the switching groups over the last twenty test items of the test.

Item #	45 minutes			60 minutes		
	All	EPL	ESL	All	EPL	ESL
20	.00	.00	.00	.00	.00	.00
21	.00	.01	.02	.00	.00	.01
22	.00	.01	.03	.00	.01	.01
23	.00	.01	.03	.02	.01	.02
26	.06	.06	.09	.07	.05	.10
27	.08	.08	.11	.08	.06	.11
28	.08	.08	.11	.08	.06	.11
29	.09	.09	.11	.08	.06	.11
30	.09	.09	.12	.09	.06	.11
31	.09	.09	.12	.09	.06	.11
32	.10	.09	.12	.09	.07	.12
33	.10	.10	.12	.09	.07	.12
34	.16	.16	.21	.14	.11	.19
35	.17	.17	.22	.14	.11	.19
36	.17	.17	.22	.14	.11	.19
37	.25	.26	.25	.19	.16	.23
38	.29	.30	.27	.22	.20	.26
39	.29	.31	.27	.22	.20	.27
40	.30	.32	.27	.23	.21	.27

Table 2: The cumulative proportion of EPL and ESL examinees in the strategy switching groups for each time condition.

Looking particularly at the distributions after item 36, the last item associated with the penultimate reading passage, we see, as expected, that both testing time and linguistic competence seem to affect the switching strategy. In general, those examinees who had 60 minutes to complete the test had a somewhat lower rate of switching to a random response strategy, 19% for the 60 minute group versus 25% for the shorter, 45 minute time condition. The difference is greater among EPL students, 26% v.s. 16%. Contrary to our expectation, the shortened testing time seemed not to affect strategy switching for the ESL examinees. Across groups, however, the organization and structure of the reading test is captured well by the cumulative proportions in each of the strategy switching categories presented in Table 2. We see, for example, that there are two points where the cumulative proportions show marked changes, at the 34th and 37th items. These two items correspond nicely to the structure of the test itself, since both are the first items in the testlets associated with the test's last two reading passages.

5. Discussion

The HYBRID model accomplished the objective that was set, namely, to account for a certain type of speededness. Clearly the model is limited to a specific type of speededness and the real world of examinee response strategy use may prove itself to be quite different. For example, examinees may randomly respond to the items in the first part of the test because of an unfamiliar content area. The current model cannot isolate such an event, and its occurrence would result in incorrect item parameters estimates. More importantly, taking the

estimates of a switched subgroup at face value may also be misleading for the following reasons. If items located in the latter part of a test are more difficult - and tests are often designed this way - many of the less skilled examinees' responses on such items may not be very different from random responses even if they had ample time to study the questions. The model would classify such examinees with nearly equal probability as either among those who switched to random responses or as those best characterized by lower proficiency estimates. Thus the resulting response patterns could not be interpreted solely within the context of test speededness. Further evaluation of the bivariate posterior distribution of ability and switch point based on the each individual response pattern would be required to reach conclusions about those uninformative distributions.

One may choose a definition of test speededness in terms of the effects on the performances of the examinee who falls in a restricted ability range, e.g., above average, above passing score, and so on. One of the objectives of monitoring and controlling for the speededness of a test is to minimize the effects of speededness on ability estimates, especially for those who are capable but slow, by reducing the absolute number of non-responses. The analyses presented earlier indicate the HYBRID models utility for doing exactly that.

The HYBRID model permits estimates of the ability and the switch points for each examinee using the current model. Thus it allows us to determine the relevancy of test speededness. Moreover, the ability estimates directly out of this model should not be considered as equivalent to ordinary IRT estimates of the ability. Testing involves not only examinees responses to test items, but also examinees' understanding of scoring methods being utilized. Full disclosure of administration procedures should be provided to examinees. Through this approach, it is quite important now to define our view on the familiar notions of power and speed of the test.

Application of the current model may not be limited to traditional testing conditions. Without much further modification, the model could be used for the sequential administration of tests by computers. With some modification, the model could be applied to a more general mixture of IRT and latent class models.

It was remarked earlier that this HYBRID model for the speededness of the test is only one of many possible extensions to accommodate qualitative interaction between categorical characteristics of examinees and items. For example this particular extension incorporated the information dealing with the location of items. However, the model is quite flexible and could incorporate opportunity to learn information as well. For example, let us suppose that a mathematics test including geometry items was administered to two different kinds of examinees, one who has taken geometry classes and the other who has never taken such a class. Although the examinees may perform equally well on the items not related to geometry, most likely their performance on the geometry items would differ drastically. Multiple sets of parameters in the context of the differential item difficulties are of the primary interests of Rost's (1990) Mixed Rasch Model and Mislevy & Verhelst's (1990) IRT for multiple strategies. The important application of the measurement model is that of applying two distinct models to the responses of an individual.

The data used in this study do not allow us to investigate the stability of item parameter estimates from a single test administration. Future study is needed to investigate the feasibility of using the HYBRID model to gain more useful information about items during field trials. It may be, for example the well known phenomena of the item parameter instability between the field test and the operational tests test is partly due to test speededness. Such instability can be identified and corrected by the HYBRID model.

Evaluating the fit of a measurement model is a crucial aspect of the process of modeling test taking behavior; yet the standard has not been set, in light of the fact that such a well-established measurement model as the IRT itself still awaits a standard method to evaluate model fit. The HYBRID model also has not established a standard method for model selection and testing. In the meantime, an information criterion AIC by Akaike (1985, 1987) or the direct likelihood method by Aitkin (1989) may be used to evaluate the fit of multiple non-nested measurement models.

The conditions which affect the appropriateness and accuracy of modeling response data with the HYBRID model are; the number of items, and the difficulty of items in the latter portion of the test. Forty or more items per examinee is recommended for use with the model. Indeed, if a majority of items at the end of the test are very difficult, a bivariate posterior distribution of switching points-by-ability indicate that the posterior variance is quite large near the end of the test.

There are three aspects to the potential contribution that the model can make in the field of testing. First, the model estimates IRT parameters with less bias, thus minimizing the impact of the speededness of the test. Second, the model provides a measure which can be used to set test length. Finally, the model reduces bias of the ability estimation for subpopulations when the proportion affected by speededness is different among subpopulations. The interaction of test speededness with subpopulations defined by demographic variable may explain some cases of differential item functioning, i.e., some portion of the DIF may be attributable to the differential test speededness.

References

- Aitkin, M. (1989). Direct likelihood inference. unpublished manuscript, Tel Aviv University.
- Akaike, H. (1985). Prediction and entropy. In A. C. Atkinson and S. E. Fienberg (Eds.), *A Celebration of Statistics* (pp. 1-24). New York: Springer-Verlag.
- Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, 52, 317-332.
- Bejar, I. I. (1985). *Test speededness under number-right scoring: an analysis of the Test of English as a Foreign Language*. Research Report (RR-85-11), Princeton, NJ: Educational Testing Service.
- Bock, R. D., and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika*, 46, 443-459.
- Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society (Series B)*, 39,1-38.
- Mislevy, R. J. and Verhelst, N. (1990). Modeling item response when different subjects employ different solution strategies, *Psychometrika*, 55, 195-215.
- Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, 14, 3, 271-282.
- Secolsky, C. (1989). *Accounting for random responding at end of the test in assessing speededness on the Test of English as a Foreign Language*. TOEFL Research Report 30, Research Report (RR-89-11), Princeton, NJ: Educational Testing Service.
- Yamamoto, K. (1987). A model that combines IRT and latent class models, Unpublished doctoral dissertation, University of Illinois, Champaign-Urbana.
- Yamamoto, K. (1989). HYBRID model of IRT and latent class models. ETS research report series (RR-89-41), Princeton, NJ: Educational Testing Service.
- Yamamoto, K. (1990). HYBIL: a computer program to estimate HYBRID model parameters. Princeton, NJ: Educational Testing Service.