DOCUMENT RESUME

ED 395 035 TM 025 055

AUTHOR Yamamoto, Kentaro
TITLE Estimating the Effects of Test Length and Test Time on Parameter Estimation Using the HYBRID Model.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-95-2; TOEFL-TR-10
PUB DATE Mar 95
NOTE 50p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Ability; Estimation (Mathematics); *Item Response Theory; Multiple Choice Tests; *Responses; *Statistical Bias; *Test Length; *Timed Tests
IDENTIFIERS *HYBRID Model; Item Parameters; Missing Data; *Speededness (Tests); Test of English as a Foreign Language

ABSTRACT

The traditional indicator of test speededness, missing responses, clearly indicates a lack of time to respond (thereby indicating the speededness of the test), but it is inadequate for evaluating speededness in a multiple-choice test scored as number correct, and it underestimates test speededness. Conventional item response theory (IRT) parameter estimation ignores the mixture of random response during calibration; consequently, estimated parameters are biased. The HYBRID model (K. Yamamoto, 1989) was extended (Yamamoto, 1990) to characterize when each examinee switches from an ability-based response strategy to a strategy of responding randomly. The model has allowed the evaluation of test speededness by estimating the proportions of examinees who switch strategies at any possible point in the test. The estimated IRT parameters based on the HYBRID model were more accurate than the ordinary IRT-only analysis. With the extended HYBRID model applied to the data taken from an experimental form of the Test of English as a Foreign Language it was found that: (1) test length had a small impact on the proportion of examinees affected by test speededness; (2) a greater proportion of examinees was affected by speededness of a test with a 50-minute time limit than a test with a 55- or 60-minute time limit; and (3) the difference in the proportions of examinees affected by the speededness of tests under 55- and 60-minute time limits was small. However, nearly 20% of examinees were affected by speededness after completing 80% of the test, so that the last 20% of responses of 20% of examinees did not represent their true ability. (Contains 4 tables, 16 figures, and 18 references.) (Author/SLD)

TOEFL®

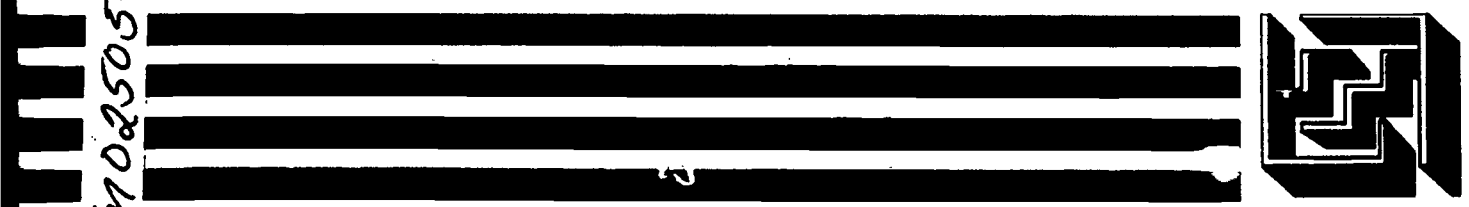
March 1995

Technical Report

TR-10

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

Kentaro Yamamoto



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

Kentaro Yamamoto

Educational Testing Service
Princeton, New Jersey

RR-95-2

Educational Testing Service is an Equal Opportunity/Affirmative Action Employer.

Abstract

When individuals perform tasks, they differ from each other not only in their ability to perform the tasks correctly, but also in their speed. Even though the traditional indicator of test speededness, missing responses, clearly indicates a lack of time to respond (thereby indicating the speededness of the test), it is inadequate for evaluating speededness in a multiple-choice test scored as number correct and underestimates test speededness. Conventional IRT parameter estimation ignores the mixture of random responses during calibration; consequently, estimated parameters are biased.

The HYBRID model (Yamamoto, 1989) was extended (Yamamoto, 1990) to characterize when each examinee switches from an ability-based response strategy to a strategy of responding randomly. The model has allowed us to evaluate test speededness by estimating the proportions of examinees who switch strategies at any possible point in the test. The estimated IRT parameters based on the HYBRID model were more accurate than the ordinary IRT-only analysis.

With the extended HYBRID model applied to the data taken from an experimental form of TOEFL®, we found that 1) the test length had a small impact on the proportion of the examinees affected by the speededness of the test, 2) a greater proportion of examinees were affected by speededness of a test with a 50-minute time limit than a test with a 55- or 60-minute time limit, and 3) the difference in the proportions of examinees affected by speededness of tests under 55- and 60-minute time limits was small. However, nearly 20 percent of the examinees were affected by speededness after completing 80 percent of the test. In other words, the last 20 percent of the responses of 20 percent of the examinees did not represent their true ability.

The Test of English as a Foreign Language was developed in 1963 by the National Council on the Testing of English as a Foreign Language. The Council was formed through the cooperative effort of more than 30 public and private organizations concerned with testing the English proficiency of nonnative speakers of the language applying for admission to institutions in the United States. In 1965, Educational Testing Service (ETS) and the College Board assumed joint responsibility for the program. In 1973, a cooperative arrangement for the operation of the program was entered into by ETS, the College Board, and the Graduate Record Examinations (GRE®) Board. The membership of the College Board is composed of schools, colleges, school systems, and educational associations; GRE Board members are associated with graduate education.

ETS administers the TOEFL program under the general direction of a Policy Council that was established by, and is affiliated with, the sponsoring organizations. Members of the Policy Council represent the College Board, the GRE Board, and such institutions and agencies as graduate schools of business, junior and community colleges, nonprofit educational exchange agencies, and agencies of the United States government.

A continuing program of research related to the TOEFL test is carried out under the direction of the TOEFL Research Committee. Its six members include representatives of the Policy Council, the TOEFL Committee of Examiners, and distinguished English as a second language specialists from the academic community. The Committee meets twice yearly to review and approve proposals for test-related research and to set guidelines for the entire scope of the TOEFL research program. Members of the Research Committee serve three-year terms at the invitation of the Policy Council; the chair of the committee serves on the Policy Council.

Because the studies are specific to the test and the testing program, most of the actual research is conducted by ETS staff rather than by outside researchers. Many projects require the cooperation of other institutions, however, particularly those with programs in the teaching of English as a foreign or second language. Representatives of such programs who are interested in participating in or conducting TOEFL-related research are invited to contact the TOEFL program office. All TOEFL research projects must undergo appropriate ETS review to ascertain that data confidentiality will be protected.

Current (1994-95) members of the TOEFL Research Committee are:

Paul Angelis	Southern Illinois University at Carbondale
James Dean Brown	University of Hawaii
Carol Chapelle	Iowa State University
Joan Jamieson	Northern Arizona University
Linda Schinke-Llano	Millikin University
John Upshur (Chair)	Concordia University

Acknowledgments

Special acknowledgment is due to the following for their substantial contribution to the success of this project:

Drew Gitomer, Gordon Hale, Robert Kantor, Jacqueline Ross, Linda Tang, and Rebecca Zwick for review of the manuscript.

Neal Thomas and Walter Way for their suggestions on the interpretation of the results.

Liz Brophy for her secretarial assistance.

The TOEFL Research Committee for their comments and support.

Table of Contents

List of Tables

List of Figures

List of Figures (continued)

Introduction

The speed of performing a task is one of the more noticeable ways in which individuals differ from each other, in addition to the ability to perform a task correctly. However, traditional methods of assessing test speededness are limited to analyses of distributions of missing responses. They tend to focus on consecutively missing responses at the end of a test. Even though some missing responses clearly indicate a lack of time to respond (thereby indicating the speededness of the test), more often it is not clear why some examinees do not respond to the items. Moreover, analysis of missing responses is inadequate in evaluating the speededness of a multiple-choice test in which the result is summarized, particularly in terms of the total number of correct responses. With such a test, the most sensible strategy for an examinee who is running out of time would be to fill in any remaining responses hoping to increase the total number of correct responses by chance. Disregarding such random responses made by more informed (or test-wise) examinees surely underestimates the actual speededness of the test. Therefore, in order to obtain a more accurate assessment of speededness, it is necessary to evaluate not only missing responses, but also random responses. This study was initiated to investigate the applicability of a new measurement model for TOEFL data analysis, as well as to make a specific time limit recommendation.

Failure to model speededness directly can also 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 test's time limit. 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 how an examinee performs within a time limit. In this paper, the notion of speededness is extended to include deterioration of responses due to lack of time.

Two previous studies (Bejar, 1985; Secolsky, 1989) relate to the current paper. Both concluded that current ETS criteria for evaluating test speededness, i.e., a test is not speeded if virtually all examinees reach the first 75 percent of the items and at least 80 percent of the examinees complete the test, are not applicable for "rights only" scored tests. Bejar based his study on the notion that on the difficult items, lower-ability examinees would perform better than predicted due to random or patterned responding. He proposed an index that compares the observed performance on the most difficult items of the test to performance predicted by the Item Response Theory (IRT) model for these items. For several ability levels, an index analogous to chi-square was calculated based on the observed proportion correct and expected proportion correct for the IRT model. However, the method is circular, as Bejar noted, because the IRT parameters were estimated on the suspected speeded data. So, the estimated item parameters are biased due to the speededness of the test. In addition, the index does not differentiate between whether the misfit of the three parameter IRT model or the speededness of the test increased the index. In fact, Bejar found misfit in both extreme

ability regions, very high and very low regions, where the IRT model parameters are least accurately estimated. Because of the shortcomings noted earlier, Bejar's index was unable to detect speededness in the most populous ability region.

Secolsky (1989) examined two exploratory techniques based on regression analysis. Both techniques are based on the idea that under unspeeded conditions scores in the beginning portion and end portion of the test should be highly correlated, i.e., the score in the beginning portion should predict the end portion score successfully. For a speeded test, the relationship between the scores for these two portions should be substantially weaker. He concluded that the test examined in the study was "slightly speeded" since the observed scores of a few examinees on the last four to six items were significantly different from the scores predicted from the first four to six items. However, any regression method based on four to six items is less than reliable. Since neither technique he used isolated error of classification due to uncertainty, any speededness found in the test may have been caused by lack of reliability alone. In addition, both of Secolsky's techniques examined speededness at a particular (somewhat arbitrary) point in a test; this restriction ignores individual differences in response speed.

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 subpopulations, whatever the subpopulation definitions may be.

The effects of different time limits on essay questions were examined by Hale (1992). He found that shorter time limits on an essay question affected all examinees at various ability ranges almost uniformly, by lowering quality of writing. Hence, altering time limits was inconsequential to the relative standing of examinees. However, his study may not generalize to a test consisting of largely multiple-choice questions. The Test of Written English (TWE®) includes only one essay question, so every examinee who is motivated can at least start the essay question. There is no variability in terms of the number of essay items attempted by examinees; therefore, test speededness can only be observed in terms of the quality of responses. Hale's study is irrelevant to a test which includes a number of multiple-choice questions. In a typical multiple-choice test, speededness can manifest itself in reduced quality of the performance as well as in number of items attempted. Therefore, both speededness and biased parameter estimation in scale linking are issues to be investigated.

Almost all procedures that link scales over time are based on the commonality of the items, the populations, or both. When the proportion of subpopulations in the total population changes, as it did with the TOEFL test population, where the composition of major language groups has changed in the last 10 years, the common population method for linking scales over time cannot be used. Common-item linking requires that item parameters be invariant over time, but it has been observed that the location of items in the test strongly affects item parameter estimates in ways that are consistent with speededness effects.

Recent developments in item-response modeling – namely, the extended HYBRID model (Yamamoto, 1990) for speededness of the test – address these problems. The HYBRID model changes the question concerning speededness from "Is a test speeded?" to "How speeded is a test?" The HYBILm computer program was written to estimate HYBRID model parameters; it estimates the proportion of examinees who switch to a guessing strategy at each item sequentially in the test. When examinees switch to a guessing strategy in the middle of a test, the probabilities of their 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. This model expands the notion of the speededness of a test to include changes in conditional probabilities in addition to the lack of responses. This model-based approach enables us to examine the effects of speededness on the estimated item parameters as well. It is particularly useful for a test based largely on the difficulty of items, from easy to difficult.

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. There are two relevant psychometric models that include multiple item-response models in a limited way. The HYBRID model by Yamamoto (1989) describes the mixture of examinees whose responses can be characterized by either an IRT-based (ordered classes) or a latent class-based (unordered class) item-response model. The Mixed Strategies model by Mislevy and Verhelst (1990) can be summarized as a correspondence between the choice of a strategy employed by an examinee and sets of IRT parameters that best describe response probabilities under each strategy. The statistical characteristics of an item may differ greatly depending upon the strategy employed. Both the HYBRID model and the Mixed Strategies model use multiple item-response models to describe the behavior of all examinees, but still only one model per examinee is posited for all of the items throughout the test.

Research into test-taking behavior, however, highlights the psychological importance of how examinees employ and switch solution strategies (Kyllonen, Lohman, and Snow, 1984.) Such behavior can be described by a mixture of several distinct item-response models for each examinee. The HYBRID model proposes to incorporate

cognitive structure into psychometric models. This model includes latent classes (guessing classes) that represent unique cognitive processes, in addition to the IRT model, which represents the proficiency of examinees in a continuum. Since a solution strategy can be thought of as an example of a cognitive structure, the basic HYBRID model hints at the possibility of different solution strategies, yet assumes that a given examinee uses the same strategy throughout the test. The original HYBRID model has been applied to real datasets by Mislevy and Verhelst (1990) and Gitomer and Yamamoto (1989). Both applications exposed the cognitively relevant structure of the data beyond the capability of the ordinary IRT model. The HYBRID model is extended here to the case of strategy switching; i.e., a subset of an examinee's responses is best described by a latent class, while IRT is most appropriate for the rest of the responses.

Developing psychometric models that incorporate strategy switching is important for three reasons: 1) to characterize the examinees' strategy usage when it is salient, 2) to detect extraneous strategy influences in estimated model parameters, and 3) to provide an opportunity to incorporate partial knowledge of latent classes. The modified HYBRID model attempts to provide a type of qualitative evaluation of the knowledge that examinees possess, and it accomplishes the objective in a limited way by relying more on the qualitative aspect of the examinee's cognitive characteristics-by-items interaction. One specific interaction will be examined closely in this paper – the interaction of the test-taking speed of examinees with the location of items in the test. This effect occurs in speeded tests in which the number right are scored. Specifically, it occurs when examinees are running out of time and switch from a strategy of thoughtful response to a strategy of patterned or random response. In other words, the propensity to make a correct response for an item is conditional not only on ability but also on test-taking speed. Standard IRT cannot handle this phenomenon and can yield misleading inferences about the proficiencies of the examinees and the properties of the test items.[1]

The Model

When tests are speeded, and when the scoring is based on the number of correct responses as is true with the GRE General Test and the TOEFL test, patterned responses are observed frequently at the end of the test, e.g., option one may be selected for the last few items. This occurs when slow test takers run out of time near the end of the test and start responding randomly to the remaining items. Unless the algorithm of the patterned responses is obvious, such as in the previous example, it is quite difficult to determine whether a segment of responses is patterned.

---

[1] Waller (1976) modeled guessing behavior by having two types of function for different ranges of ability, namely, one-parameter IRT and a flat ICC if conditional probability given ability is less than a critical value.

Built into the modified HYBRID model are the assumptions that:

1) under a patterned response strategy, the conditional probability of a correct response is independent of one's ability

2) each examinee's response to an item can be characterized either by an IRT model of a particular form or a patterned response model

3) conditional independence holds, given item parameter and an examinee's ability and strategy

Currently, the model limits the strategy switch to occur once, from a strategy that can be approximated by an IRT model, to a patterned response strategy. In many large-scale assessments and achievement tests, the great majority of omitted responses are found at the end of the tests, so the model is designed to be most responsive to capturing such a case. However, this switch-only-once assumption is not very rigid, because when one uses a probabilistic model, minor deviation from this modeling structure can be tolerated, especially when the switch occurs earlier in the test. As noted earlier, the model is probabilistic and considers probabilities of all possible switching points for every examinee; hence, deviation from the assumption has a minimal effect on the overall probability structure. For example, if an examinee skips the fourth of five passages and goes on to the fifth (and perhaps most difficult) passage, the model would find nearly equal probability of switching anywhere between the fourth 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 due to the fact that 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 fourth passage, but correctly answered questions in the fifth passage, the model would find two main probable switching locations -- one at the beginning of the fourth passage and the other at the end of the exam. The following function expresses the propensity of correct response on an item i based on the above assumptions. The notation k indicates the last item answered under the IRT model and 1.7 is most often used for D. Where $m = -1$ when $i \leq k$ and $m = 0$ when $i > k$, x_i is a dichotomous response (0 = wrong, 1 = right) on item i, β_i represents item parameters (a, b), θ is ability, c_i is the expected proportion correct under patterned response strategy.

$$P(x_i = 1 | \theta, \beta_i, k) = (1 + \exp(-Da_i(\theta - b_i)))^m c_i^{m+1} \quad (1)$$

The above function can be understood as: the 2PL IRT model holds until switch point, and then a constant conditional probability for random responses holds for the remainder of the items.

Parameter Estimation

Equation 1 gives the conditional probability of the response x_i given θ, item parameters β_i, and strategy switch point k. Equation 2 gives the likelihood of observing a response vector x_j given θ_j for a subject j who switched the strategy at item k_j.

$$P(\boldsymbol{x}_j|\theta_j,\ \mathbf{B},\ k_j) = \prod_{i=1}^{k_j} P(\theta_j,\ \beta_i)^{x_{ij}}\, Q(\theta_j,\ \beta_i)^{1-x_{ij}} \prod_{i=k_j+1}^{I} c_i^{x_{ij}}\,(1 - c_i)^{1-x_{ij}} \tag{2}$$

(Notice that for those examinees who did not switch the response strategy, the likelihood is identical to the IRT-only model.)

The marginal probability of observing x_j given model parameters **B** is,

$$P(\boldsymbol{x}_j|\boldsymbol{B}) = \sum_{k}^{I} \int_{\theta} P(\boldsymbol{x}_j|\theta,\ \boldsymbol{B},\ k) f(\theta|k)\, d\theta f(k) \tag{3}$$

where $f(\theta|k)$ is the conditional probability of θ given a switch point k, and $f(k)$ is the marginal distribution of the strategy-switching population.

The joint likelihood of parameters given the observed response matrix $\mathbf{X} = (x_1, x_2, \ldots, x_J)$ from a total of J examinees is,

$$L(\boldsymbol{B}|\boldsymbol{X}) = \prod_{j=1}^{J} P(\boldsymbol{x}_j|\boldsymbol{B}) \tag{4}$$

The IRT item parameters can be estimated to maximize the above marginalized likelihood function using an iterative method such as the Newton-Raphson (N-R) method. The N-R method can be described as $P^{n+1} = P^n - D_2^{-1} * D_1$, where P^{n+1} is a vector of parameters updated from P^n by a certain amount designated by the function D_2 (matrix of second derivatives) and D_1 (vector of first derivatives). However, D_2 can be quite large and the diagonals need not be zero. Consequently, straight application of the N-R method would be too great a computational burden. Bock and Aitkin (1981) advanced the idea of using the EM algorithm developed by Dempster, Laird, and Rubin (1977) with probit analysis inner cycles in the area of IRT parameter estimation by replacing continuous theta with discrete theta points, chosen as convenient quadrature points for the integration. With respect to u, a model parameter including either an item parameter or a population density, the first derivative of the log-likelihood of the above function can be expressed as,

$$\frac{\partial \ln L(\boldsymbol{B}|\boldsymbol{X})}{\partial u} \doteq \sum_{j=1}^{J} \sum_{k=1}^{I} \int_{\theta} \frac{\partial P(x_j|\theta,\ \boldsymbol{B},\ k)}{\partial u} \frac{f(\theta|k)\, f(k)}{P(x_i|\boldsymbol{B})}\, d\theta \tag{5}$$

Followed by the application of the empirical Bayes method and approximation of integration by summation denoted by q-quadrature points and $A(\theta_q|k)$ as defined as conditional weights approximating $f(\theta_q|k)$, the above equation for an item parameter u_i can be written as,

$$\frac{\partial \ln L}{\partial u_i} \doteq \sum_{k} \sum_{q} \frac{A(\theta_q|k)}{P_{ik}(\theta_q)Q_{ik}(\theta_q)} \frac{\partial P_{ik}(\theta_q)}{\partial u_i} \sum_{j=1}^{J} \left[x_{ij} - P_{ik}(\theta_q)\right] f(k)\, P_i(\theta_q|\boldsymbol{x}_j,\ k) \tag{6}$$

The right side of the above equation can be rewritten as follows, since x_{ij} can be either 1 or 0.

$$\sum_{k} \sum_{q} \frac{1}{P_{ik}(\theta_q)\, Q_{ik}(\theta_q)} \frac{\partial P_{ik}(\theta_q)}{\partial u_i} f(k)\, \left(R_{iqk} - P_{ik}(\theta_q)\, N_{iqk}\right) \tag{7}$$

where

$$R_{iqk} = \sum_{j} x_{ij} \frac{P(x_j|\theta_q,\ \boldsymbol{B},\ k)\, A(\theta_q|k)}{P(x_j,\ \boldsymbol{B})}$$

$$N_{iqk} = \sum_{j} \frac{P(x_j|\theta_q,\ \boldsymbol{B},\ k)\, A(\theta_q|k)}{P(x_j,\ \boldsymbol{B})}$$

and

$$\frac{\partial P_{ik}(\theta_q)}{\partial a_i} = D(\theta_q - b_i)\, P_{ik}(\theta_q)\, Q_{ik}(\theta_q)$$

$$\frac{\partial P_{ik}(\theta_q)}{\partial b_i} = -Da_i\, P_{ik}(\theta_q)\, Q_{ik}(\theta_q)$$

The matrix of second derivatives can be expressed as follows,

$$\frac{\partial^2 \ln L}{\partial a_i^2} = D^2 \sum_k \sum_q f(k)\,(\theta_q - b_i)^2\, N_{iqk} P_{ik}(\theta_q) Q_{ik}(\theta_q)$$

$$\frac{\partial^2 \ln L}{\partial b_i^2} = -b^2 \sum_k \sum_q a_i^2 N_{iqk} P_{ik}(\theta_q) Q_{ik}(\theta_q)$$

$$\frac{\partial^2 \ln L}{\partial a_i \partial b_i} = D^2 \sum_k \sum_q a_i (\theta_q - b_i)^2 N_{iqk} P_{ik}(\theta_q) Q_{ik}(\theta_q)$$

Once item parameters are estimated, estimation of an examinee's proficiency can be carried out using several existing methods, such as the maximum likelihood method (MLE), Bayes modal estimates (MAP), and expected *a posteriori* (EAP). The MLE of ability is described by Lord (1980) and MAP and EAP are both described by Bock and Aitkin (1980). I will describe EAP for the typical model. The EAP estimator for θ_j is

$$\hat{\theta}_j = \mathscr{E}(\theta_j | \boldsymbol{x}_j) \doteq \frac{\sum_k \sum_q \theta_q P(\boldsymbol{x}|\theta_q) A(\theta_q|k) f(k)}{\sum_k \sum_q P(\boldsymbol{x}_j|\theta_q) A(\theta_q|k) f(k)}$$

The variance of the EAP estimator is approximately

$$\mathrm{Var}(\hat{\theta}) \doteq \frac{\sum_k \sum_q (\theta_q - \hat{\theta}_q)^2 P(\boldsymbol{x}_j|\theta_q) A(\theta_q|k) f(k)}{\sum_k \sum_q P(\boldsymbol{x}_j|\theta_q) A(\theta_q|k) f(k)}$$

The posterior joint distribution of proficiency and switching population can be calculated as

$$P(\theta|\boldsymbol{X}, \boldsymbol{B}, k) \doteq \frac{P(\boldsymbol{X}|\theta, \boldsymbol{B}, k)}{\sum_q P(\boldsymbol{X}|\theta_q, \boldsymbol{B}, k)} = \frac{P(\boldsymbol{X}|\theta, \boldsymbol{B}, k)}{P(\boldsymbol{X}|\boldsymbol{B}, k)}$$

and

$$P(k|\boldsymbol{X}, \boldsymbol{B}) \doteq \frac{\sum_q P(\boldsymbol{X}|\theta_q, \boldsymbol{B}, k)}{\sum_k \sum_q P(\boldsymbol{X}|\theta_q, \boldsymbol{B}, k)}$$

The notion of prior distributions on the item parameters, proficiency distributions, and switch population distribution can be used during the maximization phase. For example, item parameters can be thought of as drawn from a particular distribution, and, therefore, updating parameters would be constrained to meet that particular distribution. Likewise, the proficiency distribution may be assumed as a normal distribution at each switching point including the last item. In addition $E(\theta|k)$ may be constrained to have a specific functional form in relation to the value of k.

An index to evaluate the fit of the model is always desired. For the ideal condition, G^2 can be used; however, use of this chi-square test on data when there is a sparse response pattern distribution is not appropriate. For more general cases, the evaluation of fit of even the IRT-only model remains elusive. In light of this, we should seek convergent evidence, such as a chi-square test and Akaike's an information coefficient (AIC). Although the chi-square distribution may not be exactly appropriate, the likelihood ratio of nested models is available to examine the model fit. For example, comparison of the fit of two models, such as a 1PL IRT versus a 2PL IRT model, can be made by examining the improvement in the log-likelihood, taking into account the number of degrees of freedom expended. When the competing modes are not nested, the aforementioned log-likelihood test is even less appropriate. In such a case, in place of the log-likelihood test, AIC can be used. The AIC is defined as -2*log-likelihood+2*df.

Method

The parameters of the current HYBRID model are IRT parameters for items and examinees' abilities, in addition to the parameters that define the distribution of the subjects who switched strategies. The original 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 c_i 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. This paper most often uses a non-informative prior distribution for the bivariate distribution of switching behavior and ability $f(\theta, k)$, except when $k = I$, $f(\theta|k = I)$ has a standard normal distribution. It is feasible to incorporate more constrained distributional forms that probably lead to more stable results. One reasonable distribution may be that each $f(\theta|k)$ is normal, and the marginal distribution of K divided by the number of items has a beta distribution. The model assumes that all tests are speeded to various degrees, in a range of hardly speeded to very speeded; these extremes could be represented by beta distribution for K.

Results

Simulation Study

Thirty pairs of item parameters a and b were generated. The a parameters were drawn from a normal distribution with a mean of 1.0 and a standard deviation of 0.4, with only values greater than 0.3 retained. The b parameters were drawn from a normal distribution with a mean of 0.0 and a standard deviation of 0.8. The relationship between a and b parameters was not built into the design, so any correlation between them is coincidental.

Together with these item parameters and 1,000 ability parameters generated from the standard normal distribution, 1,000-by-30 dichotomous responses were generated. In addition to the 1,000 IRT-only cases, 300-by-10 responses with a fixed conditional probability of correct responses of 0.2 were generated to simulate random responses. The responses of the 701st to 1,000th simulees on items 21 through 30 of the dataset (1000 IRT) were replaced with the 300-by-10 constant conditional probability responses, which we will call dataset (300-switch).

The next dataset, which we will call 300-omit, contained 300-by-10 omitted responses (coded as 3 not presented) in place of 300-by-10 random responses. The responses of the first 700 cases and responses on the first 20 items on the last 300 cases are identical to the (300-switch) dataset. This dataset contained identical information with regard to the IRT model without any contaminating responses. The rationale for having such a dataset is that unlike codes of 0 (incorrect) or 1 (correct), the response code of 3 (not presented) causes such responses to be ignored during model parameter estimation. When the IRT parameters are estimated using this dataset, the IRT item parameter estimation is based only on the portion of the data that corresponds to the IRT model; hence, the estimation error is minimized. Comparison of estimated IRT item parameters based on the competing models would be made against the estimates on this data instead of the original IRT parameters. The deviation of estimated parameters from the true parameters can be attributed to two factors: the error inherent in the simulation of the data and the usage of the wrong measurement model. For given data without replication, the best way to evaluate the appropriateness of a particular model is to compare the estimated parameters under two conditions: with a correct model and with a model to be tested. Evaluation of estimated parameters against the true parameters may not be optimal for many cases.

Three sets of model parameters were estimated: 1) ordinary 2PL IRT parameters on the 300-omit data (60 parameters to be estimated); 2) ordinary 2PL IRT parameters on the 300-switch data (60 parameters); and 3) HYBRID model parameters on the 300-switch data (60+40+19=119 parameters). Estimated item parameters and the values of -2*log-likelihood to indicate the fits of the model are presented in Table 1 with the true item parameters. The -2*log-likelihood for the IRT model on the omit data is not

comparable due to the fact that 300-by-10 responses were never included in calculating the likelihood; hence, it was not reported here. Two sets of estimated item parameters, one with the IRT-only model and the other with the HYBRID model, are plotted against the estimated parameters on the 300-omit data (Figures 1, 2, 3, and 4.) It is quite clear that HYBILm successfully eliminated the influence of the patterned response subpopulation on the estimated item parameters. The estimated item parameters of the last 10 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 the last 10 items presented in Table 2 clearly indicate the inaccuracy of parameter estimates when random responses are ignored.

Evaluation of estimated IRT parameters on the omitted data and switched data reveals that the amount of bias is positively related to the value of the slope as expected, since the slope is a ratio scale. This can be understood as the weighted sum of two item-response functions, one of the IRT and the other with flat fixed conditional probability sufficiently different from 0.5. It is clear that the impact on the resultant conditional probabilities is greater when the slope is steeper and also in one direction, assuming that the flat conditional probability is greater or less than 0.5. The RMSD for the last 10 items was .212. Deviations were all in one direction, namely, underestimation. Bias in the location parameter estimates based on the omitted data is nearly uniformly overestimated as expected, since the scale is not a ratio scale. The RMSD for the last 10 items was .558. Even though the chi-square fit statistic was not as small compared to the chi-square fit statistic of the IRT estimate on the omit dataset, the chi-square fit statistic for each item was fairly small, indicating a decent fit. The point is that the conventional model fit statistics cannot detect the presence of random responses in the dataset due to the speededness of the test. Even though the overall accuracy of the estimated item parameters was markedly improved using the HYBRID model, some biases remained. For example, both item parameters for the last item were less accurately estimated compared to the rest of items, partly due to the fact that there was 30 percent less information available in the data to estimate IRT parameters compared with the other 20 items. However, such a notion, the reduced amount of the IRT information resulting in greater error of estimation, applies to all of the last 10 items. Nonetheless, the biases are still much less than estimation based on the IRT model alone.

Accuracy of the ability parameter estimation is an equally important aspect of the application of IRT, especially if the test results are to be used to evaluate applicants using such tests as GRE and TOEFL. In Figure 5, ability estimates based on the IRT model are plotted against the estimates based on the IRT model of the 300-omit data. The IRT-only model indicates two systematic biases: 1) abilities of the first 700 simulees that did not contain any random responses are overestimated, and 2) abilities of the last 300 simulees with random responses were underestimated. Both biases were larger for those with higher ability. In Figure 6, the ability estimates based on the HYBRID model are plotted against the estimates based on the IRT model on the 300-omit data. There is no clear sign of bias for either group of simulees. The RMSDs of HYBRID model

estimates are presented in Table 2 with the IRT model estimates. It shows that the RMSD is nearly twice as large without using the HYBRID model for the last 300 simulees. It was 50 percent greater for the first 700 simulees with the IRT model.

The posterior bivariate distribution of ability and switching point based on the HYBRID estimate is plotted in Figure 7. The bivariate distributions show that the mode of switch population is very well identified at the 20th item.

The second simulation study dealt with different switching-population distributions. While the dataset for the IRT portion (1,000 IRT) was kept unchanged, the location of switching and the proportions of switching to patterned responses were changed. It was set at three groups of 100 subjects each, and each group switched after the 15th item, 20th item, or 25th item. This left the first 700 responses unchanged. For this study, only the bivariate posterior distribution was plotted in Figure 8. All other model parameters were very similar to the first simulation study. The results showed that the modes of three spike-shaped switching distributions were captured well in the posterior distribution, yet the shape of spike-like distribution was not detected accurately. Consequently, the estimated distribution was much smoother than the original distribution. This was expected due to the presence of uncorrelated measurement errors. Exact representation of spike-like distribution was only captured under no measurement-error conditions. Whenever strategy switching occurs, the location of the switching point is always more accurately estimated for higher-ability simulees than for those with lower abilities. The effect of switching strategy to the patterned responses is more drastic for those with higher abilities than for those with lower abilities, i.e., without the HYBRID model, underestimation of abilities would result for those who were affected by speededness. A more accurate estimation for those with higher abilities is an advantage of the model. It is also true that because estimated item parameters are biased when switching populations are included in the IRT-only model parameter estimation, estimates of the abilities of able simulees among those who do not switch to random responses have a positive bias.

The distribution of the switched population is fairly accurately estimated for those who switched near the middle of the simulated test. However, near the end of the test (the last few items), the cumulative distribution shows some overestimation of the proportion of the switched population. Consider, for example, the situation where the IRT ability is high because many responses on the earlier items are correct, but an error was made on the last item. Because of the difference between the expected probability of correct response and the actual response on the last item, the estimated posterior distribution of this response pattern would more likely be classified as switching to a strategy of random response at the end of the test. In order to investigate such overestimation of the proportion of a switching population, the following simulation study was performed. Using the original IRT dataset that does not include either not-reached responses or random responses, the HYBRID model parameters were estimated. The estimated switched distribution was solely due to errors. The estimated

cumulative switched proportion was 0.05, whereas if the model had estimated the proportion perfectly, it should have been 0. This type of estimation error depends upon the difficulty of items at the end as well.

TOEFL Data

In 1992, a direct assessment was conducted on the speededness of an experimental section of TOEFL. The study was described in Schedl, Thomas, and Way (1995). Tests with different numbers of items were randomly assigned to examinees during each administration of the experimental section. For reasons of operational feasibility, random assignment of test centers was used instead of random assignment of examinees [see Wild and Durso (1979), and Evans (1980)]. Each test booklet had six passages and either 48, 54, or 60 items, and there were three time limits, 50, 55, and 60 minutes, following a 3 x 3 factorial design. Three different numbers of items were constructed by first creating a 60-item test, 10 items for each of the six passages; one or two items were deleted from each passage to arrive at either 48 or 54 items.

The 48 common items administered to all examinees were used to evaluate the speededness of the test under several conditions. One item out of the 48 items exhibited a negative point biserial correlation; it was excluded from the analysis. Two separate calibrations for the model parameters were carried out, one using the IRT model alone and the other using the HYBRID model for the speededness of the test. The sample sizes and mean total number correct for each cell of the 3 x 3 design on the common 47 items are presented in Table 3. It is clear that time limits have a greater impact on the performance in terms of the total score than does the number of items. It should be noted that the constraint of the time limit was carried out site by site, while booklets with different numbers of items were spiraled within a test site, thus attaining a more random assignment for the booklets than for the time limits.

Before the model-based analyses were carried out, differences among the proportion correct of nine subpopulations for each item were examined. Figure 9 presents the results summarized by time limit, and Figure 10 presents the results summarized by number of items. If speededness affected performance on the items and if the subpopulations were equal in ability, all three lines should be nearly identical in the beginning, and the differences should increase near the end of the test. We would expect to see that examinees who had 50 minutes to complete the test would perform worse than those who had 60 minutes, and examinees with 55 minutes would perform somewhere in the middle. We would expect also that the performance deterioration would be greater when more items were in the test. Figure 9 shows that the 50-minute condition resulted in a nearly uniformly lower proportion correct than the other two time limits across all items from the beginning to the end. Neither Figure 9 nor 10 shows any notable sign of an interaction of performance and location of passages. This uniformity of performance may result from the following: 1) the three subpopulations not being

equal in ability; 2) some portion of the responses not being related to ability, hence reducing the proportion of correct responses; or 3) subjects responding differently from the outset of the test, depending upon the time limit given.

Average point biserial correlations within each of the six passages were .404, .462, .558, .447, .485, and .349. The last passage had the lowest average biserial correlation, indicating that the responses on the last passage related least to the responses for the other passages.

The 2PL model was used to estimate the IRT parameters. The number of quadrature points was 20 and the standard normal distribution was used as the prior distribution, i.e., only item parameters were estimated, two per every item. Fit of the model was evaluated using a chi-square statistic for each item while G^2 and AIC were used to assess the fit for the entire set of items. The item level chi-square statistic indicated poorer fit for the last 10 items. The average chi-square was 13.1. The number of parameters estimated for this calibration run was 94. The statistics for G^2 and AIC were 71,672 and 71,860, respectively.

HYBRID model parameters were estimated using specifications similar to the IRT-only parameter estimation: 20 quadrature points and the standard normal distribution for the IRT portion of subpopulation. A subpopulation that switched to random responses was assumed to have a normal distribution for each switching point with a different mean and standard deviation, and they were estimated simultaneously with the IRT parameters.

Fit of the item parameters indicated a far better fit than the IRT-only model, especially for the last 10 items. The average chi-square was 3.7. In fact, chi-square fit statistics for the last 10 items were similar to those of any other items. However, the G^2 and AIC were 71,515 and 71,821, respectively, only a slightly better fit than the IRT-only model. This is quite unusual from our experience with similar data in that commonly, we see a substantial increase in the fit statistics using the HYBRID model.

Figures 11 and 12 present the estimates of slope and location parameters for the two models. Comparisons of estimated item parameters revealed that both sets of estimates of slope and location parameters were nearly identically related, except for the last six items. The similarity of estimated item parameters indicates that essentially identical information can be obtained using either model. The ability estimates of the two models revealed similar results, plotted in Figure 13. Only about 15 percent had two ability estimates that differed by more than .20, about the size of the *SD* of average posterior distribution.

It is quite clear that with regard to the IRT parameter estimates on these particular TOEFL data, the two models can produce very similar results in terms of the order of slopes and difficulties. However, the precise effects of the last six items on scale linking still need to be investigated.

Even though the IRT parameters can be very similar, the distribution of a portion of the subpopulation that switched to random responses can be different under different time limits and lengths of tests. Cumulative distributions of subjects who switched to random responses are presented in Table 4 and summarized results by time limits and number of items are plotted in Figures 14 and 15. It was found that with shorter time limits a greater number of examinees appear to have switched to random responses than with longer time limits. However, the number of items does not seem to increase the size of the switching population. Exactly 10 percent of the sample switched to random responses at or before the 36th item (about 75 percent of the total items), and 25 percent of the sample switched to random responses at or before the 42nd item (about 90 percent of the total items).

The apparent insensitivity of speededness of the test to the number of items may be due to the fact that this study used a fixed number of passages. Judging from past experience, the speededness of a test can be sensitive to the number of passages; had the number of passages been varied in this study, the results might be quite different. As well, the interaction between number of items and the time limits could not be examined due to the small sample size. The effect of differences in time limits was evaluated by averaging over the three different levels of items. Under this analysis, it is clear that the 50-minute time limit is inadequate for the test because nearly a quarter of the examinees indicated that their performance was affected by speededn[illegible] of the test; i.e., their responses were indistinguishable from the random responses for the questions associated with the last passage. However, the difference between the 55-minute and the 60-minute time limits is less clear. The HYBRID analysis found a small difference between the two time limits.

With the extended HYBRID model applied to the data taken from an experimental form of TOEFL, we found that 1) the test length had a small impact on the proportion of the examinees affected by the speededness of the test, 2) a greater proportion of examinees are affected by speededness of the test with a 50-minute time limit than by tests with 55- or 60-minute time limits, and 3) the difference in proportions of examinees affected by speededness under 55- and 60-minute time limits is small.

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 reality may prove 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 be reflected by incorrectly estimated item parameters. Taking the estimates of a switched subpopulation at face value may be misleading for the following reasons. If items located in the latter part of a test are more difficult — tests are often designed that way — many of the examinees' responses on such items may not be very different from random responses. Because of the similarity of the two probabilities of correct

responses, the model would classify such examinees with nearly equal probability to be either among those who switched to random responses or those who are very poor performers best characterized by lower proficiency values. Such a subpopulation, however, should not be interpreted as answering incorrectly solely because of the speededness of the test. Further evaluation of the bivariate posterior distribution of ability and switch point based on each individual response pattern would lend support to the above conclusion by exhibiting a very uninformative distribution with regard to the switching point.

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. As a result, monitoring and controlling for the speededness of a test is important in order 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 ability of the HYBRID model to do exactly that.

The estimates of ability and the switch point are calculated for each examinee using the current model. However, unless the speed of taking a test is irrelevant, the ability estimates from this model should not be considered equivalent to ordinary IRT estimates of ability. It is quite important now to define our view on the familiar notions of power and speed of the test. In the past, the traditional IRT model did not identify the impact of test-taking speed on ability estimation. Most often, test-taking speed was ignored, and ability estimation was carried out as if all responses were made under unlimited time conditions. This assumption may be true for most but not all examinees. The new model offers an option to separate two factors, test-taking speed and propensity to make correct responses. If the main goal of a test is to assess an examinee's accuracy of performance rather than how quickly he or she makes correct responses, we may need to eliminate the impact of test speededness on the ability estimates of all examinees.

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 the application of the 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. This particular extension incorporated the information dealing with the location of items. However, the model is quite flexible to 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, those who have taken geometry classes and those who have not. 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. The uniqueness of this use of the measurement model lies in applying two distinct models, continuous and discrete, to an individual's responses.

The dataset used for the study did not allow us to investigate the stability of item parameter estimates from a single administration. Future study is needed to investigate the feasibility of using the HYBRID model to gain more useful information about items during field study. It is my conjecture that the well-known phenomenon of the item parameter instability between the field test and the real test is partly due to the speededness of the test, and such instability can be corrected by using 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, because 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.

There are three aspects of 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 that 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 conditions that affect the appropriateness and accuracy of modeling response data with the HYBRID model are as follows: 1) the number of items; and 2) the difficulty of items in the latter portion of the test. Forty or more items per examinee are recommended to use the model. If a majority of items at the end of the test are very difficult, bivariate posterior distribution of switching points by ability indicates that the posterior variance is quite large near the end of the test.

One major concern remains regarding the speededness of the test. Even under the best conditions for minimizing speededness, 48 items within 60 minutes, it seems that the proportion of examinees affected by the speededness of the test is large. Nearly 20 percent of the examinees switched to the strategy of responding randomly before 80 percent of the test was completed. Is this too large a proportion? How much is too much? Decisions regarding test length and testing time limits should be made by the program administrator. Data analysis based on the model can enable program direction to select from the competing options. Routine monitoring of the speededness of the test should be in place. It is clear that just monitoring missing responses is no longer adequate.

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. (ETS Research Report RR-85-11.) Princeton, New Jersey: 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., and Rubin, O. B. (1977). Maximum likelihood. From incomplete data via the EM algorithm. Journal of the Royal Statistical Society (Series B), 39, 1-38.

Evans, F. R. (1980). A study of the relationships among speed and power aptitude test scores, and ethnic identity. (College Board Research and Development Report RDR 80-81, No. 2, ETS Research Report RR-80-22.) Princeton, New Jersey: Educational Testing Service.

Gitomer, D., and Yamamoto, K. (1991). Performance modeling that integrates latent trait and class theory. Journal of Educational Measurement, 28-2, 173-189.

Kyllonen, P. C., Lohman, D. F., and Snow, R. E. (1984). Effects of aptitudes, strategy training, and task facets on spatial task performance. Journal of Educational Psychology, 76-1, 130-145.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, New Jersey: Earlbaum Associates.

Mislevy, R. (1986). Bayes Modal Estimation in item response models. Psychometrika, 51, 177-195.

Mislevy, R. J. and Verhelst, N. (1990). Modeling item response when different subjects employ different solution strategies. Psychometrika, 55, 195-215.

Schedl, M., Thomas, N., and Way, W. (1995). An investigation of proposed revisions to section 3 of the TOEFL test. (TOEFL Research Report 47, ETS Research Report RR-94-42.) Princeton, New Jersey: Educational Testing Service.

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, ETS Research Report RR-89-11.) Princeton, New Jersey: Educational Testing Service.

Waller, M. (1976). Estimating parameters in the Rasch model: Removing the effects of random guessing. (ETS Research Bulletin RB-76-8.) Princeton, New Jersey: Educational Testing Service.

Wild, C., and Durso, R. (1979). Effect of increased test-taking time on test scores by ethnic group, age, and sex. (GRE Board Research Report GREB No. 76-6R.) Princeton, New Jersey: Educational Testing Service.

Yamamoto, K. (1989). HYBRID model of IRT and latent class models. (ETS Research Report RR-89-41.) Princeton, New Jersey: Educational Testing Service.

Yamamoto, K. (1990). HYBILm: A computer program to estimate HYBRID model parameters. Princeton, New Jersey: Educational Testing Service.

Table 1. Original and Estimated Item Parameters in the Simulated Dataset[2]

			IRT-only model estimates				HYBRID model	
	True		300-omit		300-switch		300-switch	
Item	a	b	a	b	a	b	a	b
1	.32	-.99	1.11	-1.11	1.14	-1.09	1.15	-1.08
2	.74	-.60	.96	-.72	.98	-.72	.97	-.71
3	.56	-.09	.61	-.11	.60	-.11	.62	-.10
4	.72	.72	.56	.55	.54	.57	.57	.56
5	.84	-.44	.85	-.31	.83	-.31	.85	-.30
6	.74	-1.13	.85	-1.13	.81	-1.16	.82	-1.14
7	.42	.20	.29	.40	.28	.42	.29	.41
8	.56	.02	.38	-.03	.36	-.04	.38	-.02
9	.69	.68	.64	.95	.60	.98	.63	.96
10	.83	1.08	.88	1.26	.82	1.31	.88	1.28
11	.52	-1.15	.52	-1.03	.50	-1.05	.53	-1.01
12	.59	-.25	.47	-.34	.43	-.36	.47	-.33
13	.97	-.03	.83	.06	.80	.06	.83	.06
14	.62	-.76	.43	-1.08	.40	-1.14	.41	-1.11
15	.94	-.20	.84	-.16	.78	-.16	.86	-.16
16	.85	.10	.83	.22	.80	.22	.83	.21
17	.58	.48	.58	.32	.53	.34	.59	.32
18	.90	1.49	.72	1.72	.69	1.75	.74	1.67
19	.37	.36	.29	-.06	.30	-.06	.30	-.09
20	.33	-1.20	.38	-.83	.38	-.83	.40	-.88
21	.87	.32	1.07	.34	.83	.66	1.13	.36
22	.44	.78	.33	.99	.30	1.36	.32	.98
23	.80	-.51	.59	-.59	.48	.12	.65	-.45
24	1.21	-.22	1.33	-.14	.84	.33	1.49	-.11
25	.59	.90	.67	1.07	.50	1.50	.72	1.09
26	.51	-.77	.45	-.87	.38	-.09	.51	-.87
27	.47	.26	.34	.35	.30	.97	.33	.17
28	.44	.75	.39	.76	.37	1.19	.43	.53
29	.53	-.50	.39	-.70	.29	.06	.42	-.95
30	1.25	.69	.96	.91	.65	1.33	1.20	.79
-2*log-likelihood					34,698.7		34,350.9	
No. of Parameters					60		119	
AIC					34,818.7		34,468.9	

[2]The switched data set had two groups, 700 simulees out of 1000 did not switch strategy, and 300 simulees switched to random response of $p=0.2$ after the 20th item. Omitted data replaced random responses with 300 x 10 3s, indicating not reached.

Table 2. Root Mean Square Deviations[3] of Estimates of Model Parameters Among Two IRT-Only Calibrations and HYBRID

RMSD of slope parameter estimates for the last 10 items			
	Estimates using IRT-only		HYBRID on the 300-switch data
	the 300-omit data	the 300-switch data	
True parameters	.16	.27	.16
Estimates (IRT-only) on 300-omit	.0	.21	.10

RMSD of location parameter estimates for the last 10 items			
	Estimates using IRT-only		HYBRID on the 300-switch data
	the 300-omit data	the 300-switch data	
True parameters	.14	.59	.19
Estimates (IRT-only) on 300-omit	.0	.56	.13

MD and RMSD of ability estimates against estimates on the omit data				
Estimates (IRT-only) on the 300-omit data	Estimates (IRT-only) on the 300-switch data		HYBRID on the 300-switch data	
	MD	RMSD	MD	RMSD
first 700 cases	-.10	.12	-.05	.09
last 300 cases	.23	.37	-.01	.19
total 1000 cases	-.00	.23	-.00	.13

[3] Deviation was calculated using the following formula for every item: deviation = row estimate-column estimate. The RMSD was calculated using the above values.

Table 3. Sample Sizes and Mean Total Score for the 3 x 3 Design Cells

		Number of items			
Time Limits (min)		60	54	48	Total
60	*N*	174	180	173	527
	Mean (of 47)	28.8	27.6	27.9	28.1
55	*N*	123	120	123	366
	Mean (of 47)	28.0	28.2	26.4	27.5
50	*N*	143	149	138	430
	Mean (of 47)	24.5	26.7	24.3	25.21
Total	*N*	440	449	434	
	Mean (of 47)	27.2	27.5	26.4	

Table 4. Cumulative Distributions of Switched Population by Subgroups

Item No.	Time Limits 60			Time Limits 55			Time Limits 50			Total
	60	54	48	60	54	48	60	54	48	
28	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
29	.01	.02	.02	.02	.02	.02	.02	.02	.03	.02
30	.03	.05	.05	.04	.05	.05	.06	.05	.07	.05
31	.04	.06	.06	.05	.06	.06	.07	.06	.08	.06
32	.04	.06	.06	.05	.06	.07	.08	.07	.08	.06
33	.05	.07	.07	.06	.07	.08	.09	.07	.09	.07
34	.05	.09	.08	.07	.08	.09	.11	.09	.11	.09
35	.06	.10	.09	.08	.09	.10	.12	.10	.12	.10
36	.07	.12	.11	.10	.11	.12	.14	.11	.14	.11
37	.10	.15	.14	.13	.14	.16	.20	.15	.20	.15
38	.12	.17	.16	.15	.17	.18	.23	.17	.22	.17
39	.16	.21	.19	.18	.20	.22	.26	.20	.26	.21
40	.18	.22	.20	.20	.21	.23	.28	.22	.27	.22
41	.22	.26	.23	.23	.25	.27	.32	.25	.32	.26
42	.24	.29	.25	.26	.27	.29	.35	.28	.34	.28
43	.30	.34	.30	.31	.33	.34	.41	.33	.40	.34
44	.31	.35	.31	.33	.35	.35	.42	.34	.40	.35
45	.32	.37	.33	.34	.37	.37	.44	.36	.42	.37
46	.34	.39	.35	.37	.39	.38	.45	.37	.43	.38
47	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Figure 1

Item Slope Parameter Estimates by IRT-only Model
on 300-switch and 300-omit Data

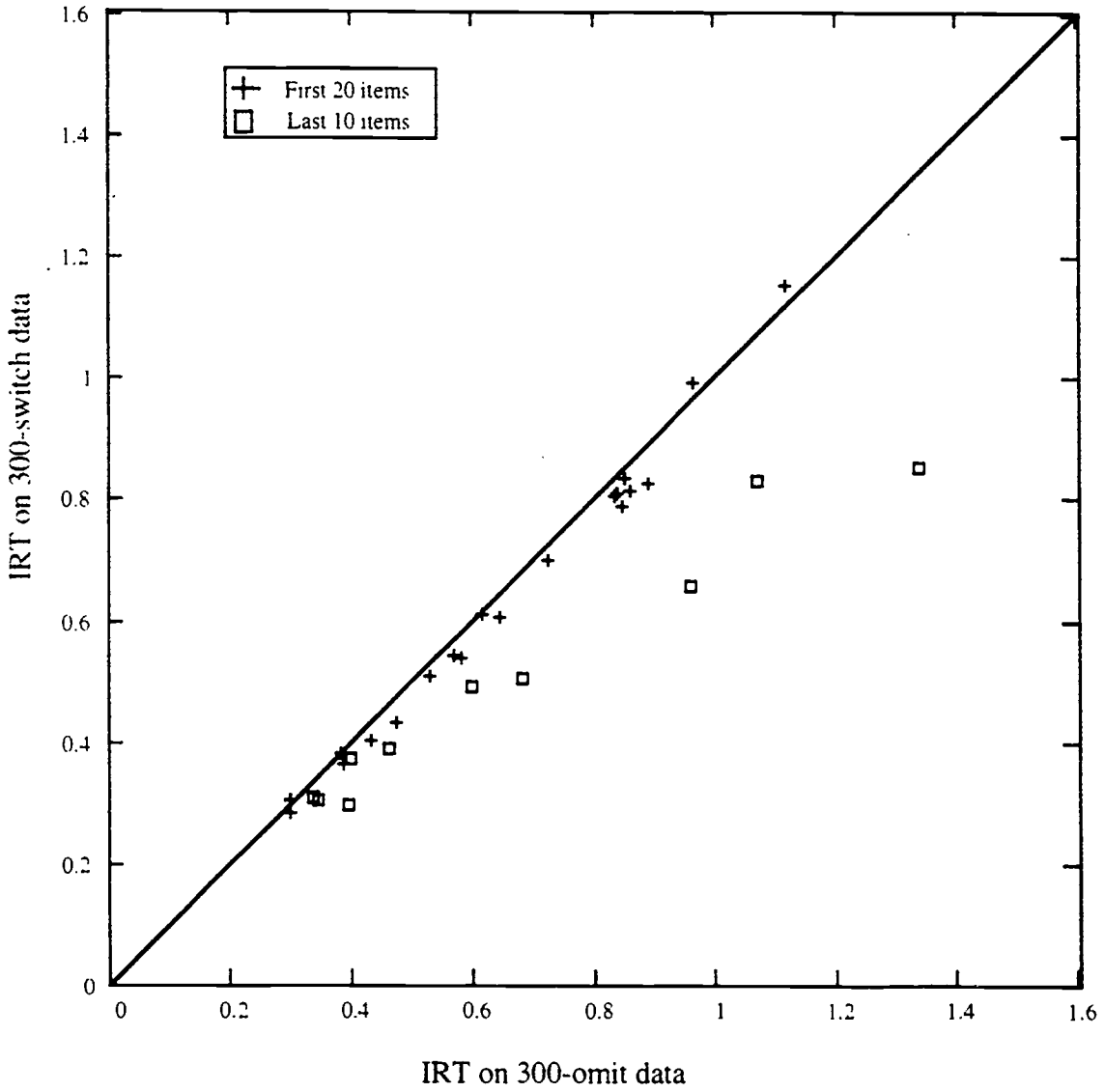


Figure 2

Item Slope Parameter Estimates by HYBRID and IRT Models
on 300-switch Data

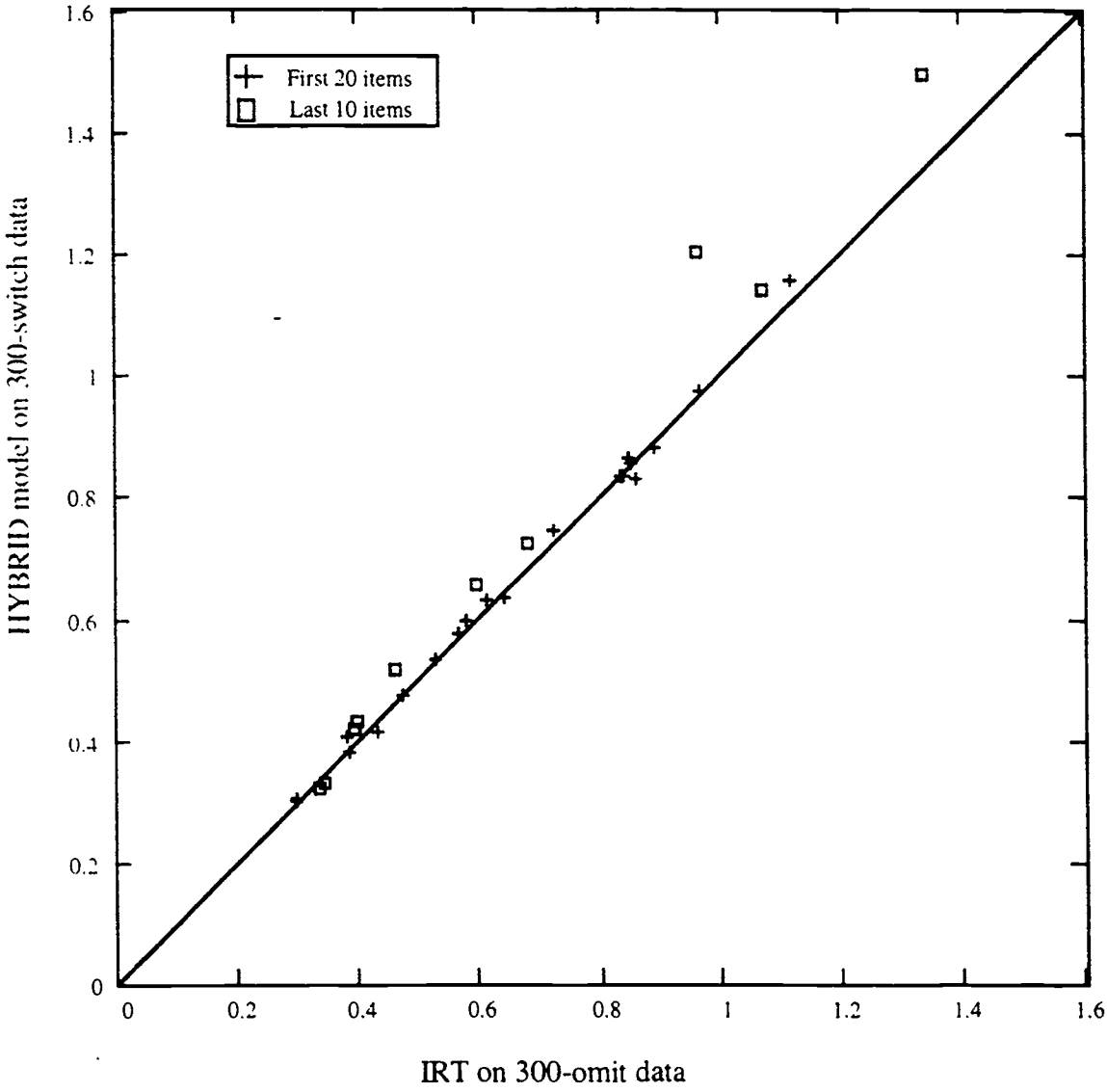


Figure 3

Item Location Parameter Estimates by IRT-only Model
on 300-switch and 300-omit Data

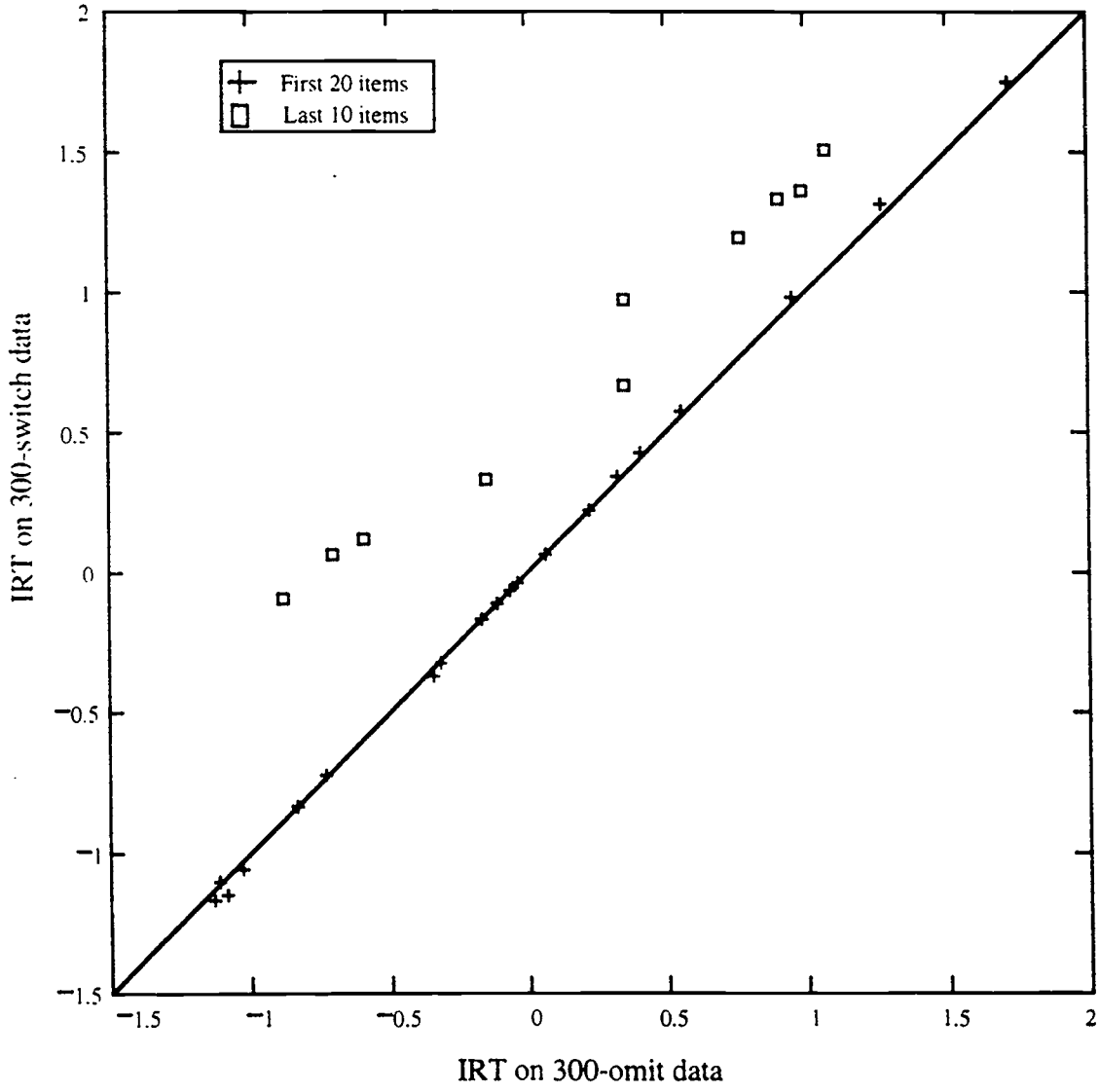


Figure 4

Item Location Parameter Estimates by HYBRID and IRT Model
on 300-switch Data

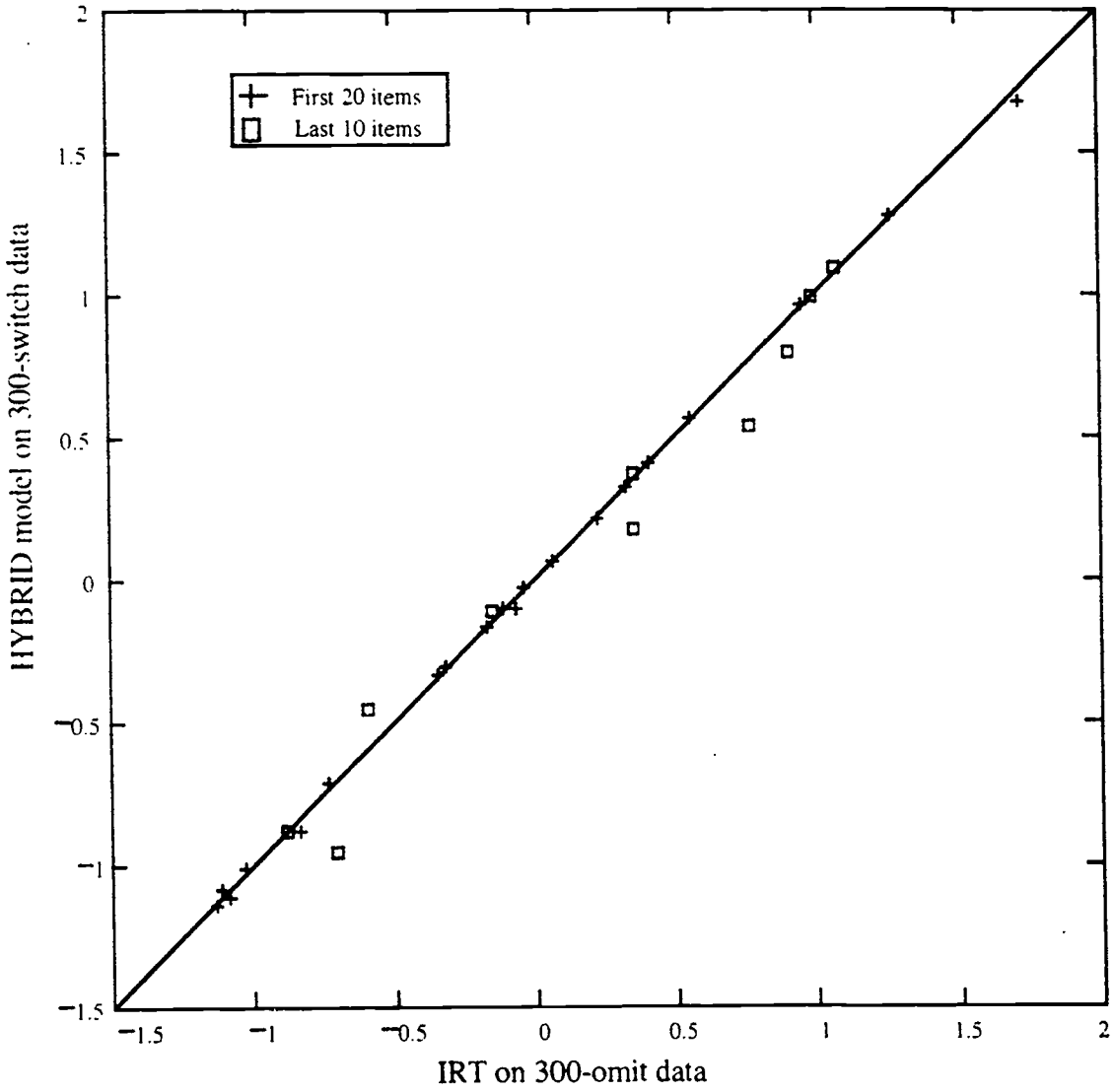


Figure 5

Ability Parameter Estimates by IRT-only Model
on 300-switch and 300-omit Data

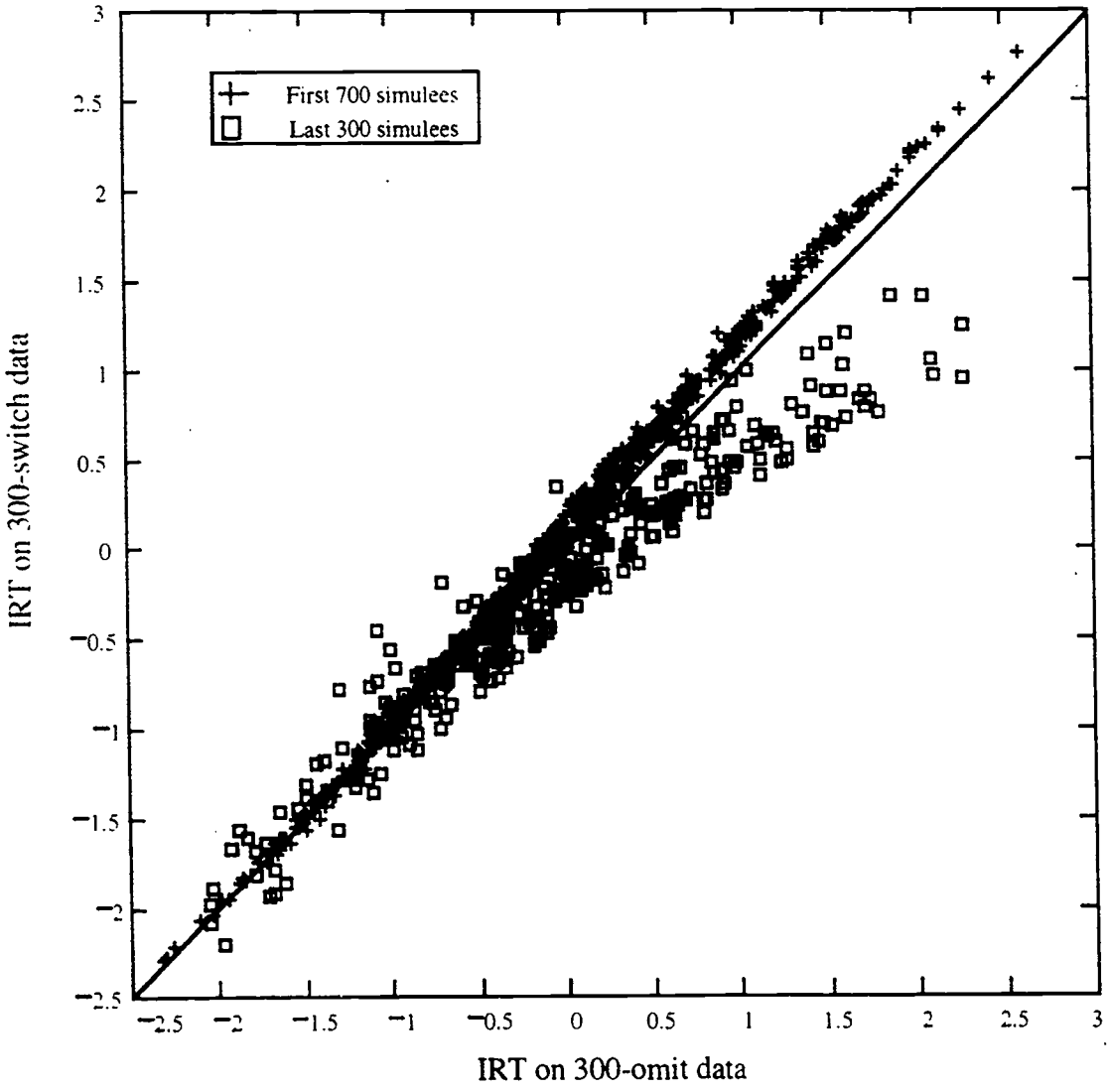


Figure 6

Ability Parameter Estimates by HYBRID and IRT Model
on 300-switch Data

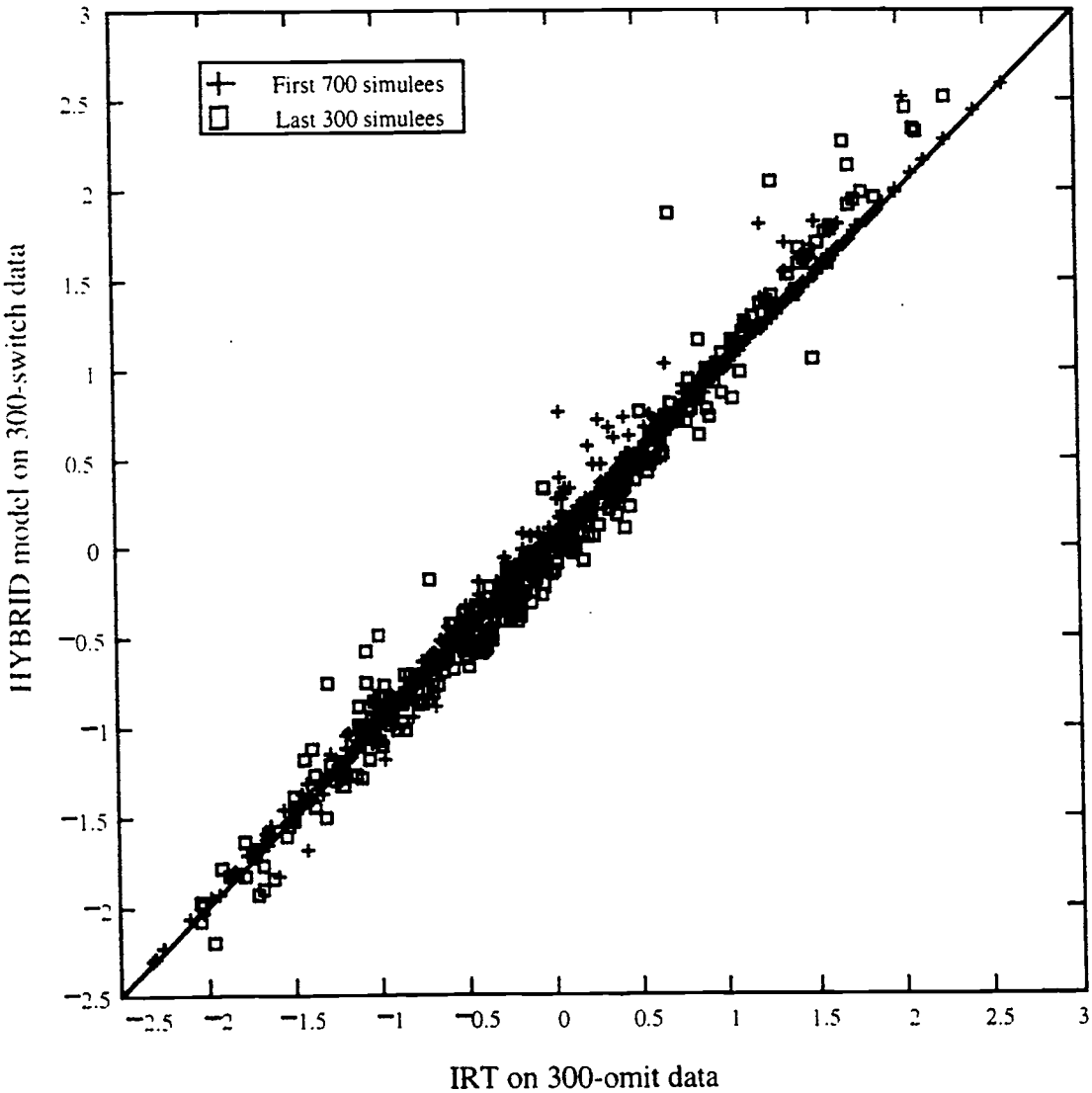


Figure 7

Estimated Posterior Distribution of Switched Population
of 300-switch Data

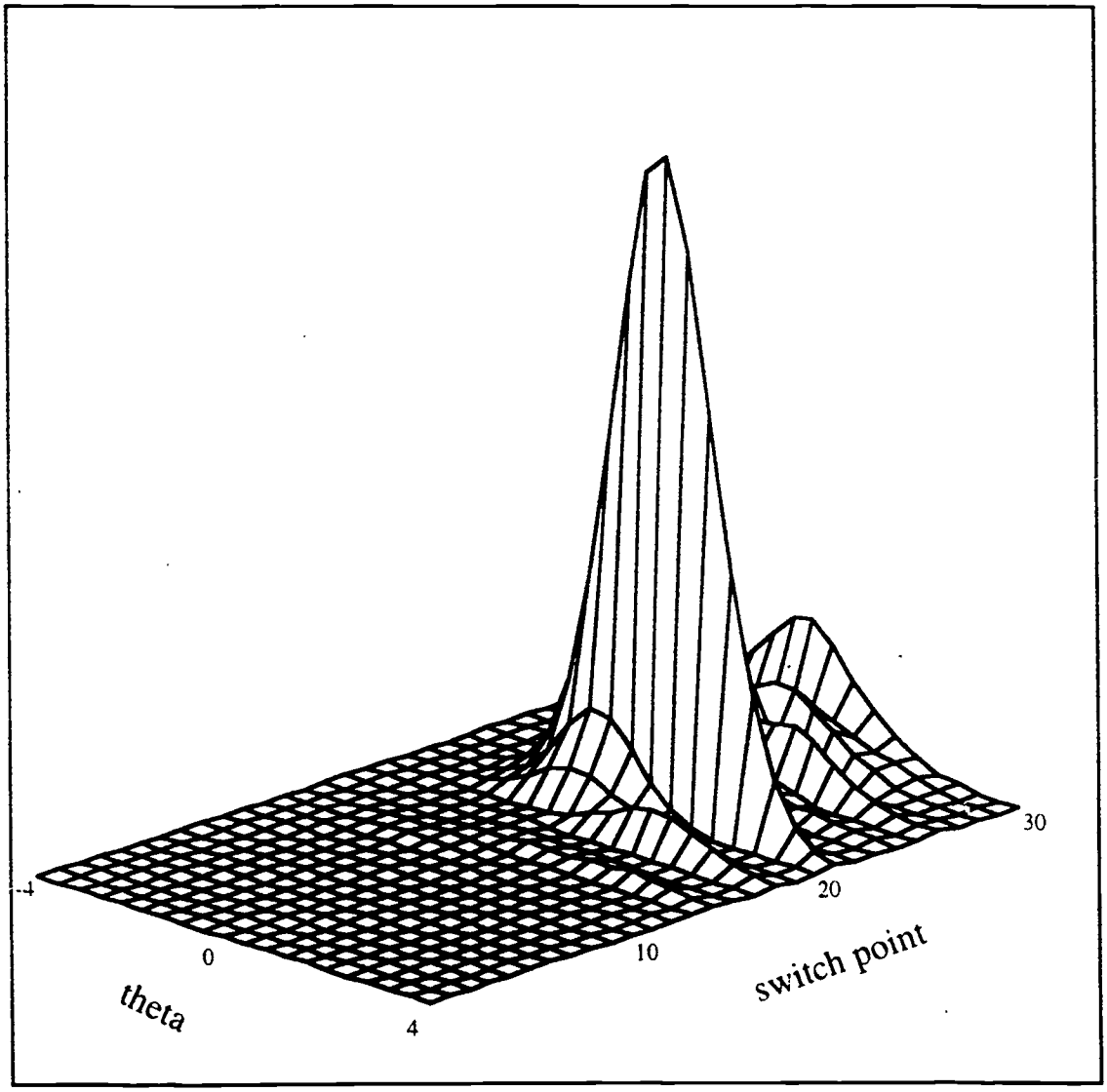
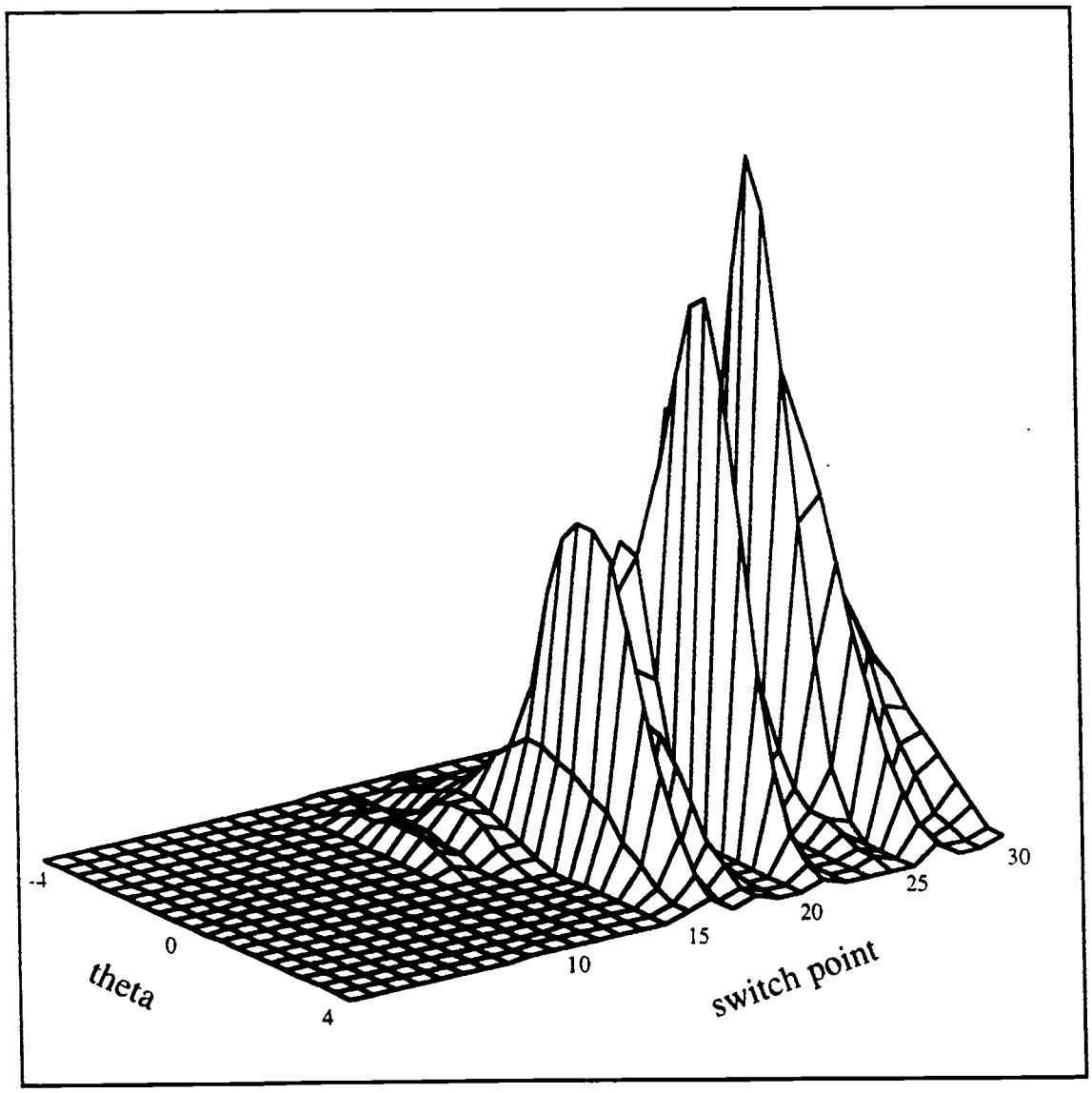


Figure 8

Estimated Posterior Distribution of Switched Population
of 300-step Data



POSTERIOR

Figure 9

Residual Item Proportion Correct of Three Time Limits Against the Mean

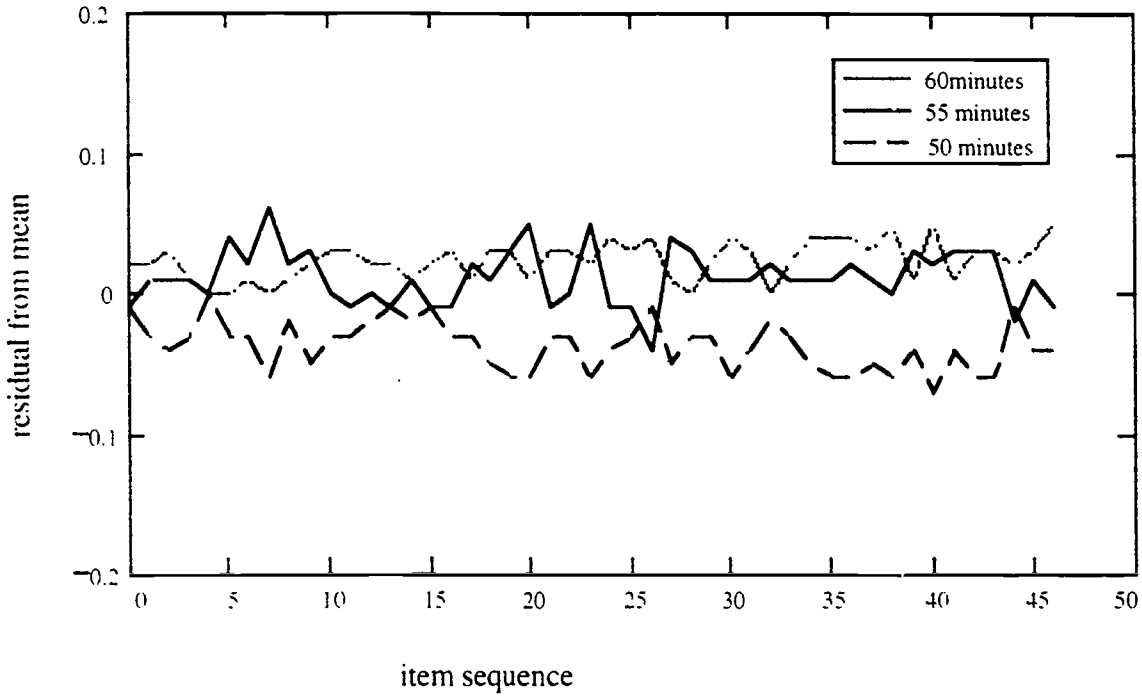


Figure 10

Residual Item Proportion Correct of Three Test Lengths Against the Mean

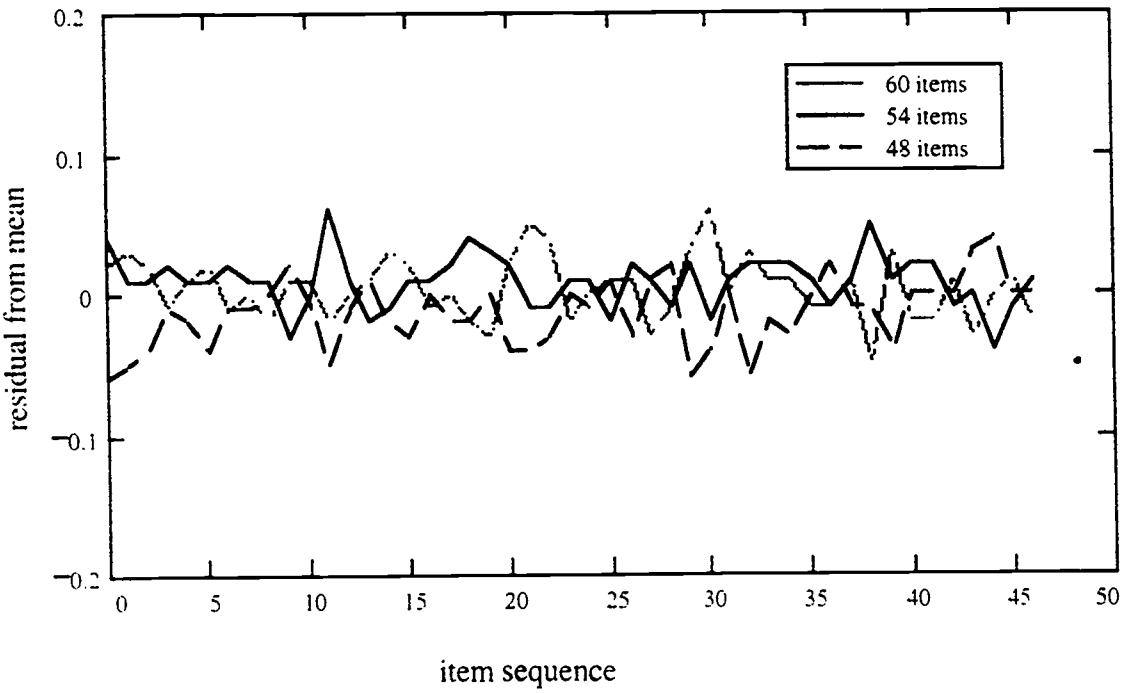


Figure 11

Slope Parameter Estimates by IRT-only Model
Against HYBRID Model

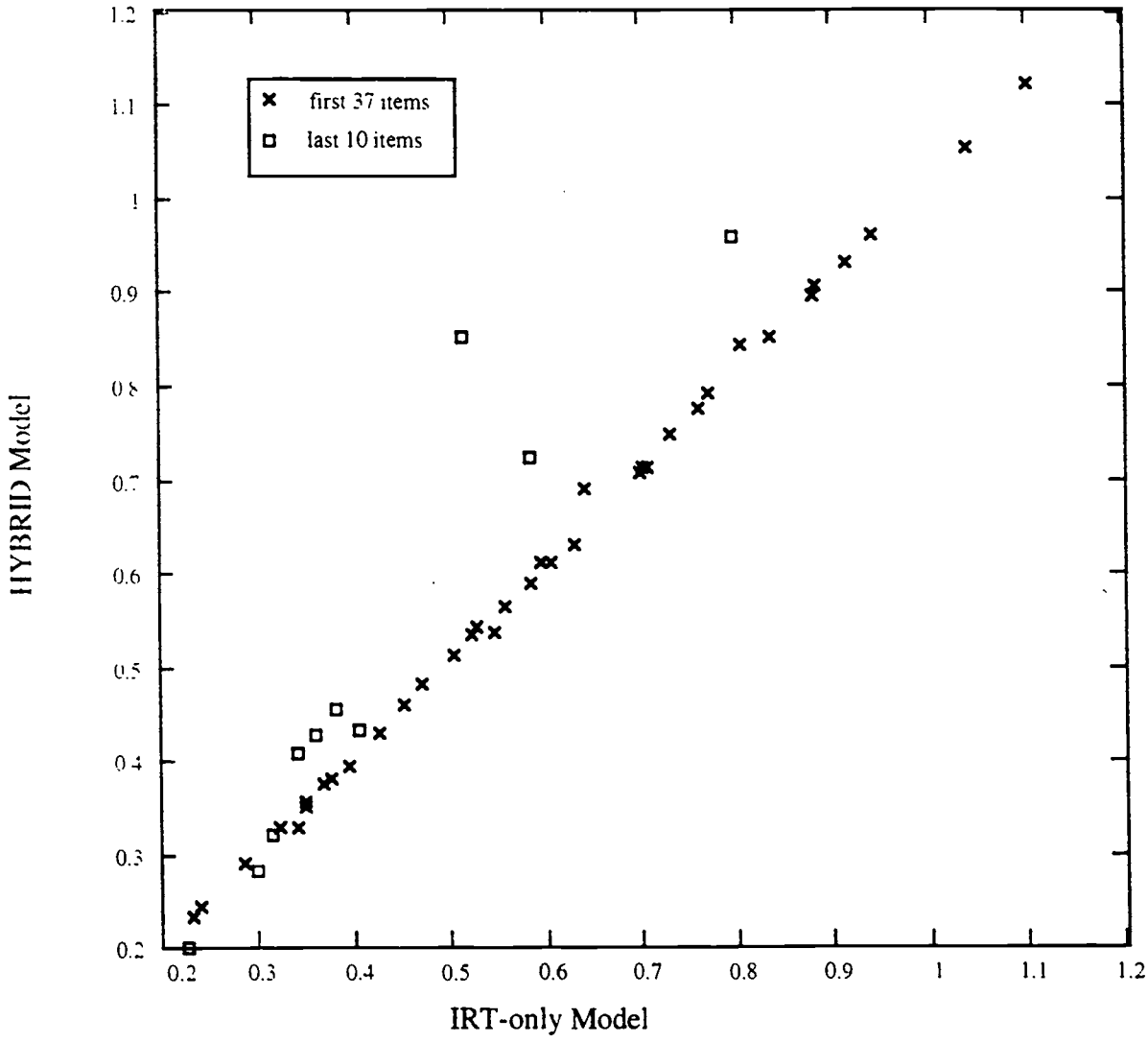


Figure 12

Location Parameter Estimates by IRT-only Model
Against HYBRID Model

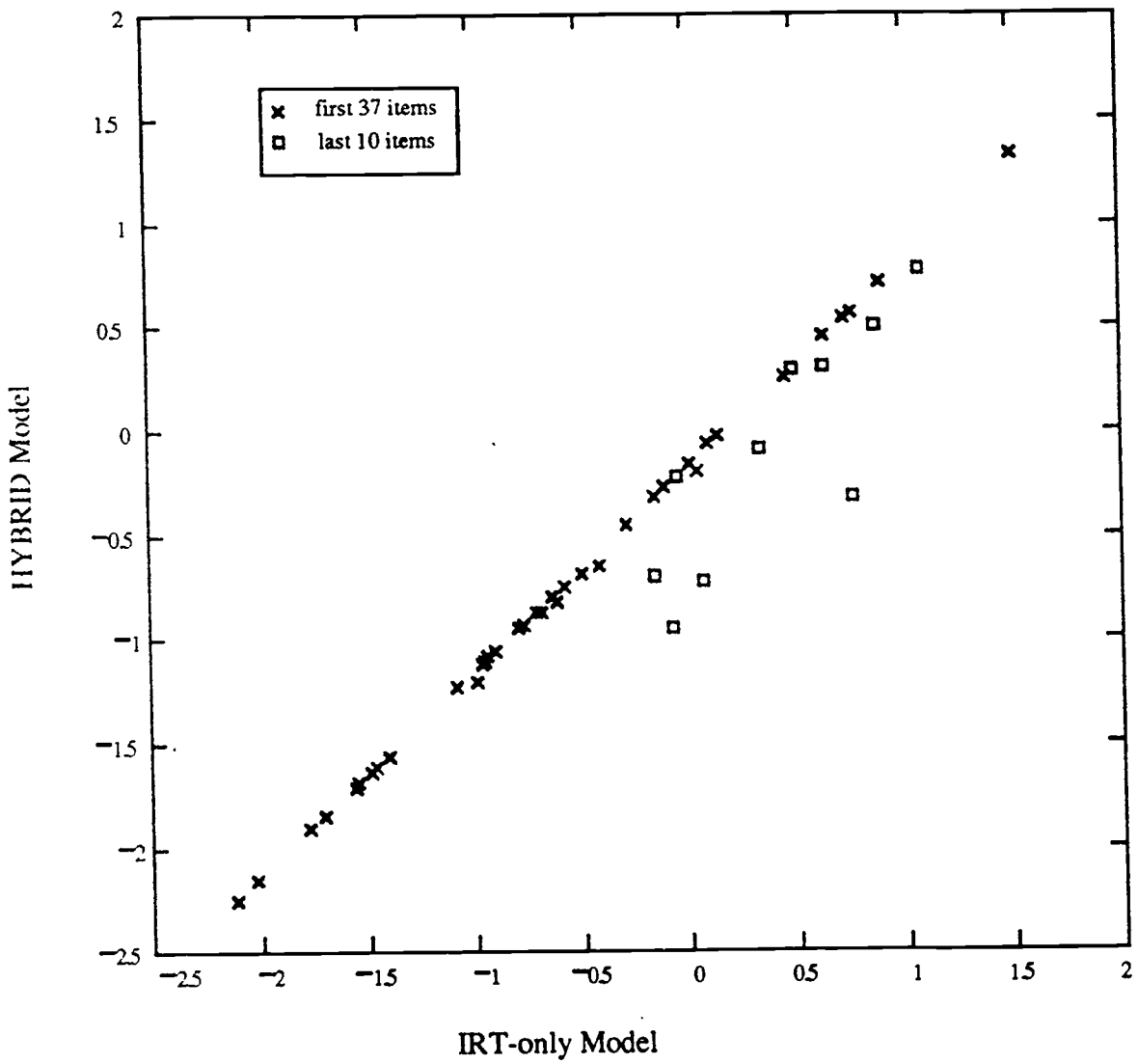


Figure 13

Ability Parameter Estimates by IRT-only Model
Against HYBRID Model

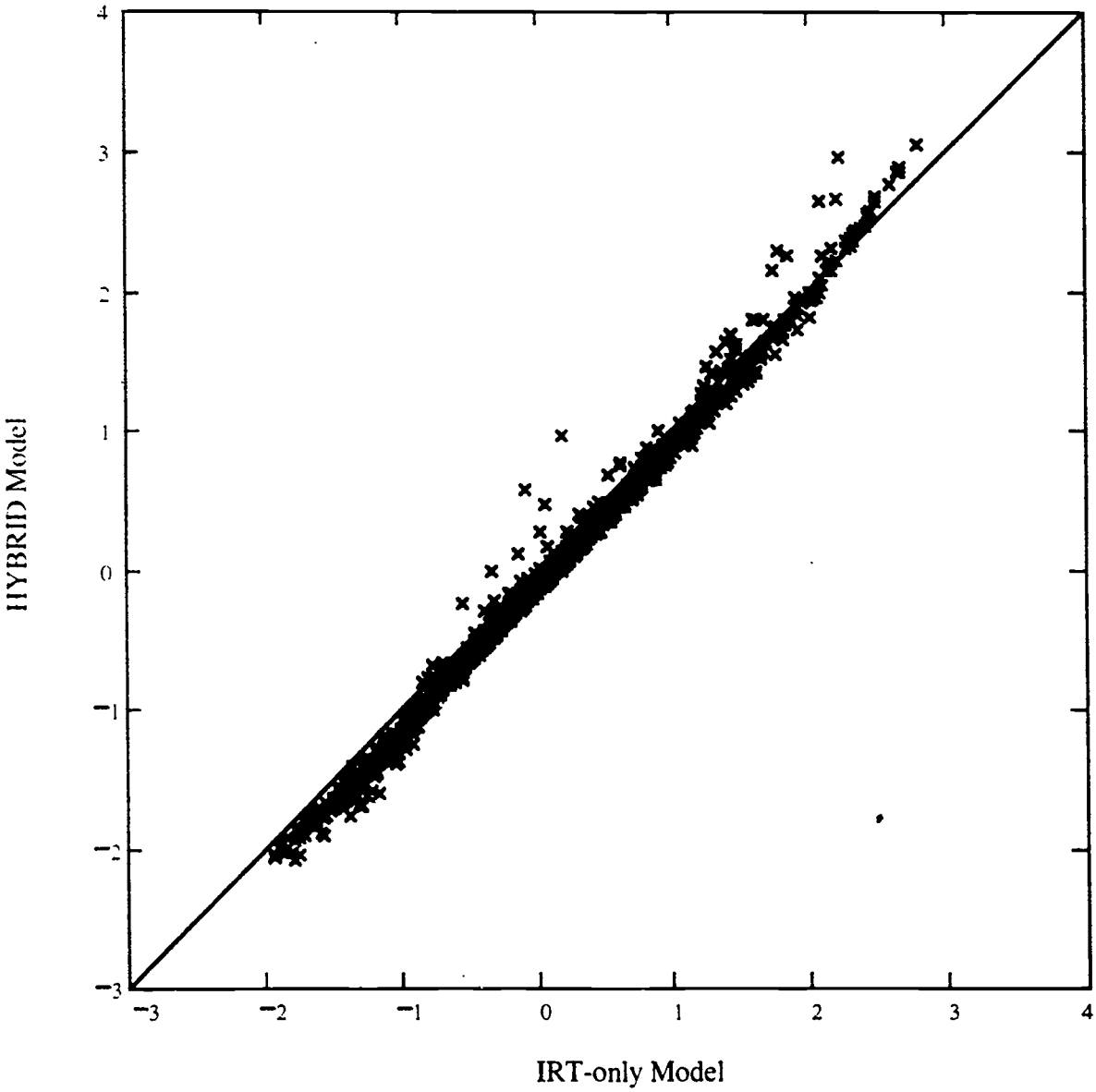


Figure 14

Cumulative Distributions of Examinees Affected by the Speededness of the Test Under Three Time Limits

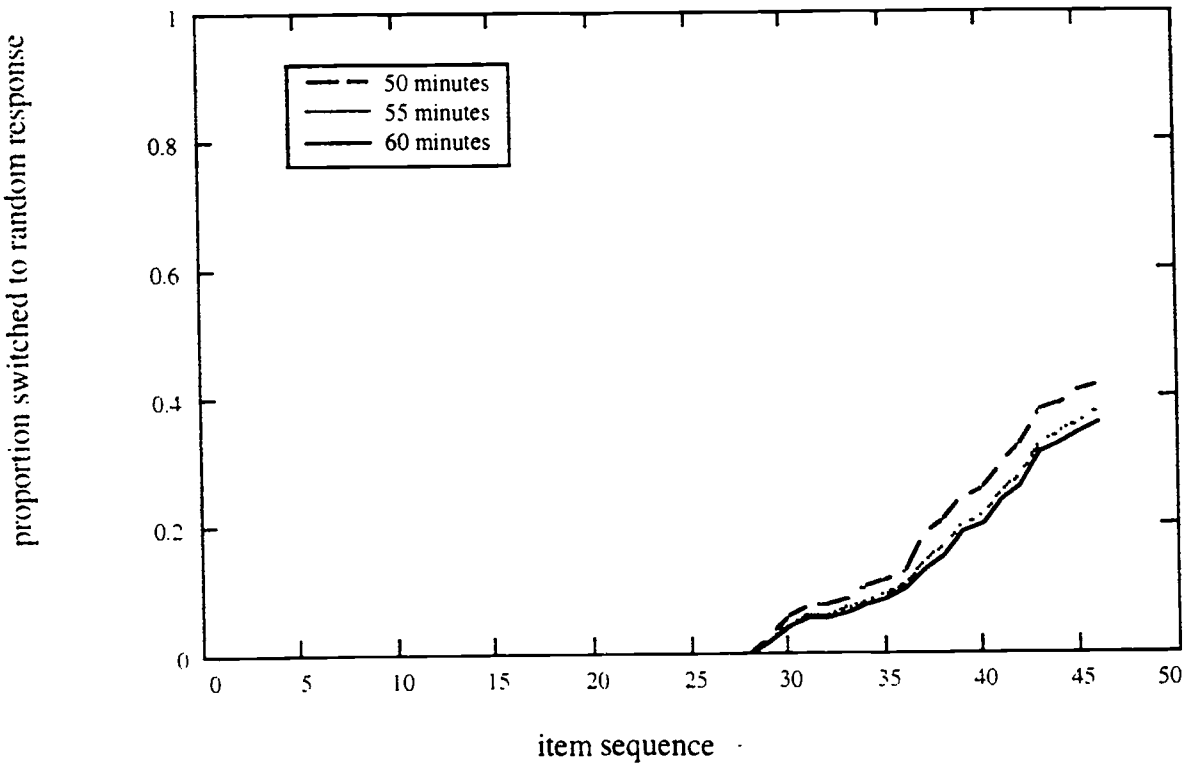


Figure 15

Cumulative Distributions of Examinees Affected by the Speededness of the Test Under Three Different Lengths of Test

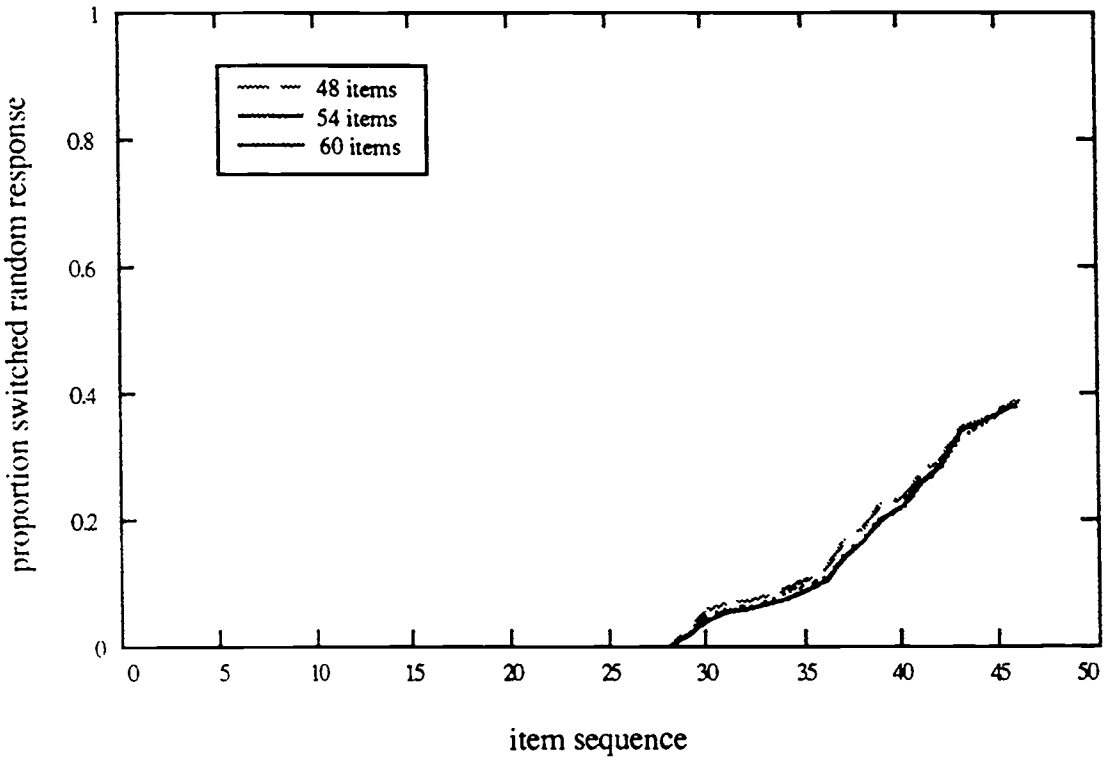
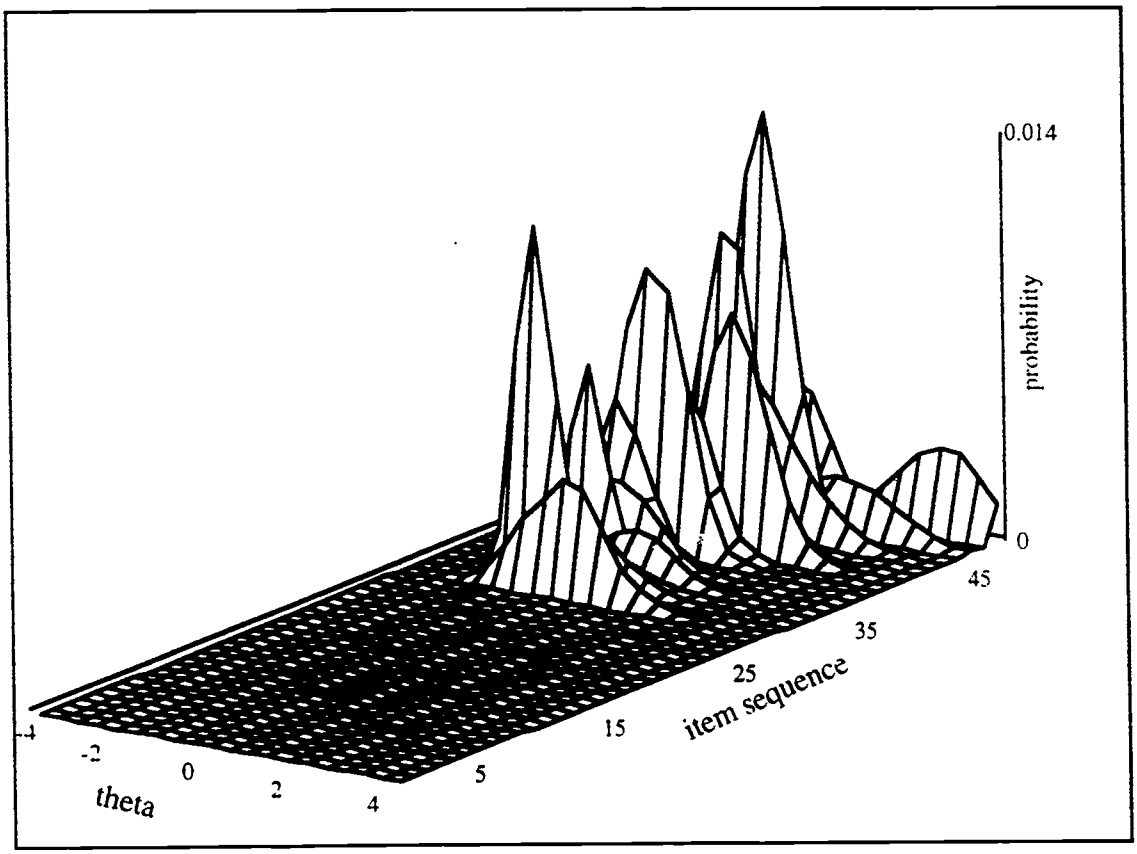


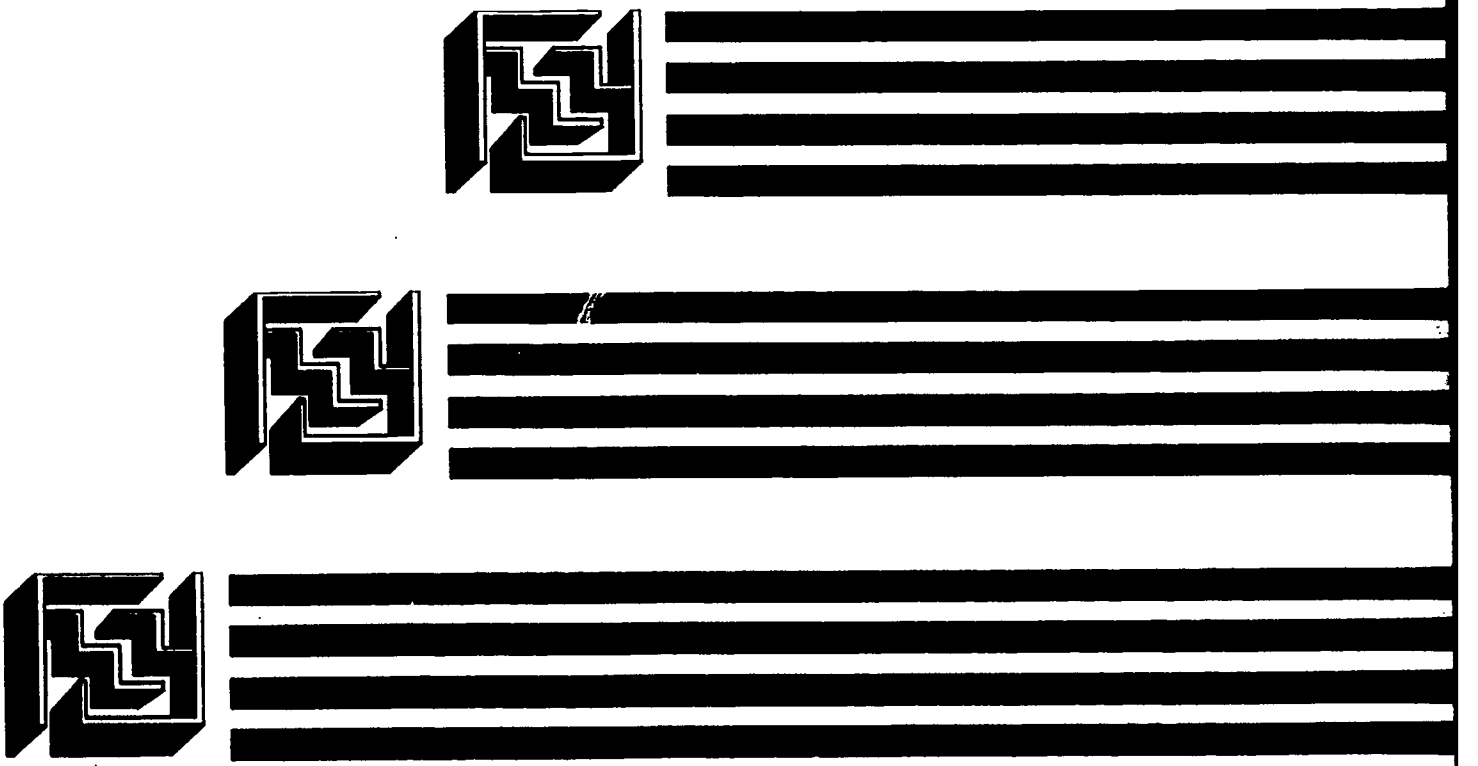
Figure 16

Estimated Posterior Distribution of Switched Population
of TOEFL Experimental Data



Posterior Distribution

TOEFL is a program of
Educational Testing Service
Princeton, New Jersey, USA



57906-12797 • Y75M.5 • 275616 • Printed in U.S.A.