DOCUMENT RESUME

ED 467 809 TM 034 351

AUTHOR Schnipke, Deborah L.

TITLE The Influence of Speededness on Item-Parameter Estimation. Law School Admission Council Computerized Testing Report. LSAC Research Report Series.
INSTITUTION Law School Admission Council, Princeton, NJ.
REPORT NO LSAC-R-96-07
PUB DATE 1999-03-00
NOTE 21p.
PUB TYPE Reports - Research (143)
EDRS PRICE EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS College Entrance Examinations; Difficulty Level; *Estimation (Mathematics); *Guessing (Tests); Law Schools; *Multiple Choice Tests; Responses; Simulation; *Timed Tests
IDENTIFIERS *Law School Admission Test; *Speededness (Tests)

ABSTRACT

When running out of time on a multiple-choice test such as the Law School Admission Test (LSAT), some test takers are likely to respond rapidly to the remaining unanswered items in an attempt to get some items right by chance. Because these responses will tend to be incorrect, the presence of "rapid-guessing behavior" could cause these items to appear to be more difficult than they really are. Using simulated date, this study found that when rapid-guessing behavior is present, items appear more difficult and less discriminating than they really are. Using response times, an attempt was made to remove responses that appeared to be the result of rapid guessing behavior. A two-state mixture model was fit to the response time distribution of each item, and responses that were more likely to come from the rapid-guessing distribution (according to the model) were removed. After removing the fast responses (rapid guesses), the item parameters, item characteristic curves, and information functions were recovered more accurately. When test data are contaminated by speededness (rapid guesses), this study shows that response times, if available, can be used to identify and remove rapid guesses and thereby recover the true item parameters more accurately. (Contains 4 tables, 8 figures, and 12 references.) (Author/SLD)

TM

The Influence of Speededness on Item-Parameter Estimation

Deborah L. Schnipke
Law School Admission Council

Law School Admission Council
Computerized Testing Report 96-07
March 1999

LAW Services

A Publication of the Law School Admission Council

The Law School Admission Council is a nonprofit corporation that provides services to the legal education community. Its members are 196 law schools in the United States and Canada.

Law School Admission Council fees, policies, and procedures relating to, but not limited to, test registration, test administration, test score reporting, misconduct and irregularities, and other matters may change without notice at any time. To remain up-to-date on Law School Admission Council policies and procedures, you may obtain a current *LSAT/LSDAS Registration and Information Book*, or you may contact our candidate service representatives.

This study is published and distributed by the Law School Admission Council (LSAC). The opinions and conclusions contained in this report are those of the author and do not necessarily reflect the position or policy of the Law School Admission Council.

Table of Contents

Executive Summary

When running out of time on a multiple-choice test such as the Law School Admission Test (LSAT), some test takers are likely to respond rapidly to the remaining unanswered items in an attempt to get some items right by chance. Because these responses will tend to be incorrect, the presence of "rapid-guessing behavior" could cause these items to appear more difficult statistically than they really are. Using data (correct/incorrect responses and response times) that were simulated to match real data, the present study found that this is indeed the case. Using the simulated response times, an attempt was made to remove responses that appeared to be the result of rapid-guessing behavior (because the responses had such short response times). A mathematical model was fit to the response time distribution of each item, and responses that were more likely, according to the model, to come from the rapid-guessing distribution (based on response times) were removed. After the fast responses (rapid guesses) were removed, estimated item difficulty was very close to the "true" item difficulty used to simulate the data. When test data are contaminated by speededness (rapid guesses), the present study shows that response times, if available, can be used to identify and remove rapid guesses and thereby recover the true item difficulty more accurately. If the LSAT is converted to a computer-delivered format, response times will be available, and we will be able to determine if rapid guesses are altering our estimates of item difficulty, and if so, to remove the rapid guesses and their influence.

Abstract

When running out of time on a multiple-choice test such as the Law School Admission Test (LSAT), some test takers are likely to respond rapidly to the remaining unanswered items in an attempt to get some items right by chance. Because these responses will tend to be incorrect, the presence of "rapid-guessing behavior" could cause these items to appear more difficult than they really are. Using simulated data, the present study found that when rapid-guessing behavior is present, items appear more difficult and less discriminating than they really are. Using response times, an attempt was made to remove responses that appeared to be the result of rapid-guessing behavior. A two-state mixture model was fit to the response time distribution of each item, and responses that were more likely to come from the rapid-guessing distribution (according to the model) were removed. After removing the fast responses (rapid guesses), the item parameters, item characteristic curves (ICCs), and information functions were recovered more accurately. When test data are contaminated by speededness (rapid guesses), the present study shows that response times, if available, can be used to identify and remove rapid guesses and thereby recover the true item parameters more accurately.

Introduction

Both classical test theory and item response theory (IRT) assume that test takers answer items after fully considering them. An incorrect answer is taken to mean that the test taker was unable to answer the item (i.e., the item was too difficult for the test taker). This is how we interpret an incorrect answer on the Law School Admission Test (LSAT). Most achievement and aptitude tests, including the LSAT, are administered with fairly strict time limits, primarily for administrative reasons. Such tests are called "speeded" or "partially speeded" because the speed at which a test taker works will affect his or her test score. On speeded tests, performance may decline because test takers run out of time. Test takers may begin to respond randomly, or after only briefly skimming the items. On such items, an incorrect answer does not necessarily mean that the item was too difficult for the test taker; the test taker may have been fully capable of answering the item correctly, given more time.

Item parameters clearly will be affected by random responding. If test takers respond randomly to items, these items will appear more difficult than they really are (e.g., Oshima, 1994). One solution to this problem for test construction and other purposes is to use a different set of item parameters for items when they are administered at the end of a separately timed test section (because these items are typically the most affected by speededness). Another solution is to hold item position constant when an item is reused (or from a pretest, when initial item parameters are typically obtained, to the operational use of the items, when the items contribute to test taker scores, as done on the LSAT).

The author would like to thank Mark D. Reckase for his helpful comments on an earlier draft.

These options may provide workable solutions in paper-and-pencil test construction and item analysis, but as testing programs consider moving to computer-administered formats, better solutions to the problem of random responding may exist. Better solutions will be especially important if the test will be given adaptively. In a computer adaptive test (CAT), an item can potentially be used in any position in the test. For a CAT to be successful, item parameters must not change depending on item position.

Although random responding is probably not the only cause of item-parameter instability, controlling for or eliminating random responding would assuredly lead to more stable and accurate parameter estimates. Yamamoto (1995) developed a model that takes random responding into account when estimating item parameters. His model (HYBRID) assumes that test takers start a test by engaging in a strategy of thoughtful response, but at each successive item, some test takers switch to a strategy of random response. In his model, thoughtful responses are modeled by IRT. Under the random response strategy, the probability of a correct response is independent of test taker ability. HYBRID does not allow for more than one switch in strategy (as might happen if a test taker responds randomly only to a certain content area or item type), and on difficult items HYBRID may not be able to distinguish random responses from thoughtful responses if items are arranged in order of difficulty. However, using such a model is undoubtedly better than ignoring the random responses completely.

If one assumes that random responses are made quickly (perhaps as time expires), response times (if available) provide an additional way to identify responses as random. Using response times, Schnipke (1995) and Schnipke and Scrams (1997) showed that "rapid guesses" were obviously present on the analytical section of a computer-administered version of the Graduate Record Examinations (GRE) General Test. The items on the last half of the test showed evidence of a second underlying distribution of very fast, primarily incorrect, responses. Schnipke and Scrams (1997) modeled the response times with a two-state mixture model and showed that the response-time distribution for each item could be described very well by a two-state model. The two states were labeled "rapid-guessing behavior" and "solution behavior," and the accuracy rates were consistent with the labels: for each item, the rapid-guessing state had a relatively flat, low accuracy rate that was near chance, and the solution state had a relatively flat, higher accuracy rate (determined by item difficulty, by definition); the accuracy rate in the area of overlap between the states was an increasing function of response time.

In the present study, simulated data[1] were used to address two research questions: How contaminated by speededness are item-parameter estimates? and What can be done about it? To address the first question (how contaminated), data (correct/incorrect responses and response times) were generated with and without rapid guesses, and the IRT parameters, item characteristic curves (ICCs), and item information functions for the two conditions were compared to each other and to the true values. To address the second question (what can be done about it), two methods were used to remove what appeared to be rapid guesses (using the modeling techniques of Schnipke and Scrams, 1996), and the item parameters, ICCs, and item information functions were recomputed and again compared to the true values.

Method

Simulated Test Takers

Ability (θ) parameters were randomly sampled from a standard normal distribution for 5,000 simulated test takers.

Item Parameters

Two types of parameters were simulated for 30 items: IRT parameters to describe the probability that a randomly chosen test taker of a given ability will answer the item correctly when engaged in solution behavior, and parameters to describe the response-time distribution for each item.

[1] A fixed-length nonadaptive design was simulated. Such data could be collected operationally whenever blocks of items (e.g., testlets) are administered together (e.g., Lewis & Sheehan's, 1990, Bayesian mastery testing design; Luecht, Nungester, & Hadadi's, 1996, computerized adaptive sequential testing design; and Reese & Schnipke's, 1996, two-stage testlet design).

IRT Parameters

The unidimensional, three-parameter logistic (3PL) IRT model was used to describe the probability that a randomly chosen test taker of a given ability would answer the item correctly when engaged in solution behavior.[2] The 3PL model is given by

$$P_i(correct \mid \theta) = c_i + \frac{1 - c_i}{1 + e^{-1.7a_i(\theta - b_i)}} \quad (1)$$

where

θ is the test taker ability parameter,

$P_i(correct \mid \theta)$ is the probability that a randomly chosen test taker with ability θ answers item i correctly,

a_i is the discrimination parameter for item i,

b_i is the difficulty parameter for item i, and

c_i is the lower asymptote parameter for item i (the probability that a test taker of very low ability will answer the item correctly).

The discrimination (a) parameters were randomly sampled from a normal distribution with a mean of 0.8 and a standard deviation of 0.3. The difficulty (b) parameters were randomly sampled from a normal distribution with a mean of 0 and a standard deviation of 1. The lower asymptote (c) parameters were randomly sampled from a uniform distribution ranging from .15 to .25 (roughly comparable to 5-option multiple-choice items). The a, b, and c parameters were generated independently. The "items" then were sorted by the difficulty (b) parameter, from lowest to highest, to imitate a multiple-choice test with increasing item difficulty. The sorted items were labeled 1 through 30, item 1 being the easiest and item 30 the most difficult. The generating parameters are shown in Table 1.

[2]Accuracy for rapid-guessing behavior was set at chance (0.2), regardless of ability, thus IRT parameters are not necessary for the rapid-guessing state.

TABLE 1

Item parameters used to generate accuracy and response times for each item

	IRT Parameters			Response Time Parameters				
Item	a	b	c	m_s	σ_s	ρ	m_G	σ_G
1	0.900	-2.310	0.166	68.813	0.620	0.000	7.389	1
2	0.558	-2.196	0.215	52.108	0.514	0.000	7.389	1
3	0.342	-2.052	0.188	85.095	0.415	0.000	7.389	1
4	0.711	-1.348	0.178	49.659	0.520	0.001	7.389	1
5	0.729	-1.166	0.217	95.915	0.479	0.008	7.389	1
6	0.965	-1.166	0.151	52.842	0.522	0.011	7.389	1
7	1.248	-0.898	0.214	48.042	0.527	0.011	7.389	1
8	1.063	-0.764	0.214	79.919	0.366	0.026	7.389	1
9	0.969	-0.514	0.233	74.007	0.452	0.039	7.389	1
10	1.452	-0.441	0.154	52.144	0.580	0.039	7.389	1
11	1.371	-0.353	0.199	48.780	0.483	0.077	7.389	1
12	0.875	-0.342	0.242	71.355	0.509	0.084	7.389	1
13	0.355	-0.280	0.250	56.677	0.540	0.114	7.389	1
14	1.368	-0.217	0.240	77.911	0.521	0.122	7.389	1
15	1.157	-0.089	0.217	52.586	0.533	0.154	7.389	1
16	1.265	0.067	0.207	47.918	0.506	0.154	7.389	1
17	0.647	0.144	0.232	70.440	0.510	0.157	7.389	1
18	0.406	0.396	0.161	85.885	0.427	0.162	7.389	1
19	0.736	0.437	0.173	60.409	0.531	0.174	7.389	1
20	0.504	0.439	0.234	56.952	0.502	0.175	7.389	1
21	0.856	0.492	0.221	78.830	0.478	0.184	7.389	1
22	0.961	0.684	0.204	46.245	0.497	0.212	7.389	1
23	0.888	0.715	0.169	63.184	0.478	0.257	7.389	1
24	0.452	0.759	0.195	38.380	0.559	0.260	7.389	1
25	1.013	0.830	0.156	46.020	0.530	0.266	7.389	1
26	1.157	1.498	0.160	53.757	0.380	0.282	7.389	1
27	0.858	1.605	0.168	81.453	0.462	0.305	7.389	1
28	0.735	1.715	0.205	53.287	0.544	0.317	7.389	1
29	1.167	1.748	0.247	53.128	0.547	0.392	7.389	1
30	0.352	2.209	0.198	48.464	0.612	0.429	7.389	1

Response Time Parameters

Response time distributions for each item were based on work by Schnipke and Scrams (1997). Schnipke and Scrams found that a two-state mixture model (e.g., Luce, 1986) described the response-time distribution for items on a speeded test very well. The two-state mixture model is given by

$$F_{Oi} = \rho_i F_{Gi} + (1 - \rho_i) F_{Si} \quad (2)$$

where

F_{Oi} is the observed response time distribution for item i,

ρ_i is the proportion of rapid guesses on item i,

F_{Gi} is the rapid-guessing response-time distribution for item i, and

F_{Si} is the solution behavior response-time distribution for item i.

Schnipke and Scrams found that rapid-guessing and solution behavior distributions could be described well by lognormal distributions. The lognormal probability density function is given by

$$f(t) = \frac{1}{\sqrt{t\sigma(2\pi)}} \exp\left\{ \frac{-[ln\,(t/m)]^2}{2\sigma^2} \right\} \quad (3)$$

where

t is the response time (RT),

m is the scale parameter (the median of the RTs), and

σ is the shape parameter (the standard deviation of the ln(RT)'s; Evans, Hastings, & Peacock, 1993).

The present study also used a mixture of lognormal distributions. Thus, to specify the response time distribution for each item, 5 parameters are required: ρ, m_G and σ_G (the subscript G indicates the rapid-guessing lognormal distribution), and m_S and σ_S (the subscript S indicates the solution behavior lognormal distribution). These 5 parameters were simulated for each of the 30 items (described next) and were used to generate the response times for each simulated test taker on each item.

Proportion of rapid guesses (ρ). The proportion of rapid guesses for each item, ρ_i, was specified such that the proportions were very similar to what Schnipke and Scrams (1997) found in empirical data. The true proportion of rapid guesses, ρ, for each item is given in Table 1. The true proportion of rapid guesses was specified as an increasing function of item number (as was item difficulty, which created a positive relationship between item difficulty and the amount of rapid-guessing behavior).

Rapid-guessing behavior distribution: m_G and σ_G. Schnipke and Scrams found that the rapid-guessing distribution could be constrained to be the same across all items and still fit the response times. They concluded that rapid-guessing behavior is essentially the same across items (i.e., it is independent of item characteristics). Thus in the present study, a common rapid-guessing distribution was used for all items with $m_G = 7.389$ seconds, and $\sigma_G = 1.0$ ln(seconds). (These values are similar to what Schnipke and Scrams found in real data.)

Solution behavior distribution: m_S and σ_S. The median and standard deviation of the solution behavior response time distribution varied across items, but were independent of item characteristics (such as difficulty, discrimination, lower asymptote, or item position). The median for each item, m_S, (which is expressed in seconds) was sampled from a lognormal distribution with median of 60.34 seconds and standard deviation of 1.0 ln(seconds), and σ_S (which is expressed in ln[seconds]) was sampled from a lognormal distribution with median of 1.65 seconds and standard deviation of 0.08 ln(seconds). A correlation of -0.3 between m_S and σ_S was built into the parameter selection as suggested by the results of Schnipke and Scrams. The resulting solution behavior parameters, given in Table 1, are similar to those found by Schnipke and Scrams.

Generating Responses and Response Times

To generate responses (correct/incorrect) and response times for each test taker item pair, it was first determined if the test taker was in the rapid-guessing or solution behavior state on the item, then accuracy and response times were generated, as described below. To determine in which state the test taker was, a random uniform number from 0 to 1 was generated. If this number was less than ρ (the proportion of rapid guesses on the item), the test taker was assigned to the rapid-guessing state for that item. Otherwise, the test taker was assigned to the solution state for the item. In this way, the proportion of test takers assigned to the rapid-guessing state was approximately ρ. The assignment to the rapid-guessing or solution behavior state was independent of ability.

If the test taker was assigned to the solution behavior state, accuracy on the item was determined from a probabilistic application of the 3PL IRT model (Equation 1) based on the test taker's simulated ability (θ) and the item's IRT parameters. Response time was randomly sampled from a lognormal distribution with the solution behavior parameters for that item (shown in Table 1).

If the test taker was assigned to the rapid-guessing state, the probability of a correct response on the item was set equal to chance (0.2), regardless of the item's IRT parameters. (A random number from 0 to 1 was generated, and if the number was less than 0.2, the response was correct; otherwise it was incorrect.)

Response time was randomly sampled from a lognormal distribution with the rapid-guessing parameters (shown in Table 1). A dataset of 0's and 1's (incorrect/correct) and response times was thus created.

A dataset was also created with $\rho = 0$ for all items (i.e., no rapid guesses). All test takers were in the solution behavior state on all items. Accuracy followed the 3PL model and response times followed the solution behavior distribution (with m_S and σ_S). By comparing the item-parameter estimates from the two datasets (with and without rapid guesses), the degree of item-parameter contamination can be determined, addressing the first research question, How contaminated by speededness are item parameter estimates?

Classifying Responses as Rapid Guesses

To address the second question, What can be done about item-parameter contamination? two methods were used to classify fast responses as rapid guesses (ignoring, of course, the true classifications used to generate the data). Responses that were classified as rapid guesses were removed during item-parameter estimation to see if the true item parameters could be recovered more accurately.

To classify responses as rapid guesses, two techniques were used: one that was probabilistic, and one that was based on a cutoff. Both techniques relied on fitting a two-state mixture model to the response times for each item. Thus before discussing the classification techniques, response-time modeling will be discussed.

Modeling Response Times

Nonlinear regression was used to fit the two-state mixture model (Equation 2) to the cumulative distribution function (CDF) of response times for each item. Specifically, the Levenberg-Marquardt algorithm, which is used by SPSS to fit unconstrained models, was used to fit the model to each CDF (described in Norušis, 1994). The underlying distributions (the rapid-guessing and solution behavior distributions) were specified as lognormal distributions (Equation 3). The rapid-guessing distribution was not constrained to be the same across items during the estimation of the two-state model parameters (although the data were generated that way). Thus 5 variables were free to vary for each item: $\hat{\rho}$ (the estimated proportion of rapid guesses), $\hat{m}_G$ and $\hat{\sigma}_G$ (the lognormal parameter estimates for the rapid-guessing distribution), and $\hat{m}_S$ and $\hat{\sigma}_S$ (the lognormal parameter estimates for the solution behavior distribution).

The true and estimated parameters are shown in Figures 1–3. As shown in Figure 1, the proportion of rapid guesses was recovered fairly well, although the estimated proportion was sometimes a little high and sometimes a little low. As shown in Figure 2, the median, m_S, of the solution behavior distribution for each item was recovered quite well, although σ_S was recovered slightly less well for some items (items 6, 17, 19, 24, 28, 29, and 30).

As shown in Figure 3, the median of the rapid-guessing behavior distribution, m_G, was recovered well for most items. The exceptions are item 6, where $\hat{m}_G$ was far too high (which corresponds to the estimated proportion of rapid guesses on item 6 being too high), and some of the items toward the beginning of the test. Likewise, σ_G for items toward the beginning of the test was not recovered well, although σ_G was recovered well for items toward the end of the test. This may be explained by the fact that the true proportion of rapid guesses was very small at the beginning of the test so the rapid-guessing distribution could not be estimated well, whereas toward the end of the test the proportion of rapid guesses was large enough to provide stable estimates of the rapid-guessing distribution.

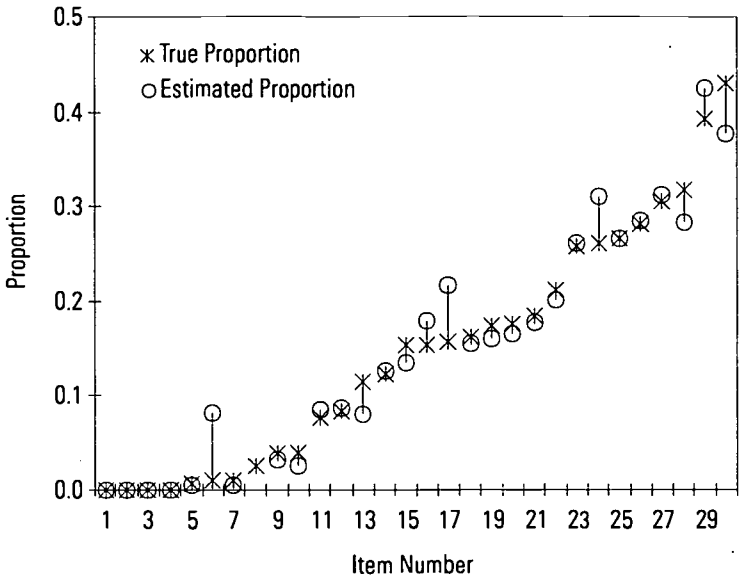

FIGURE 1. *True and estimated proportion of rapid guesses in the simulated data*

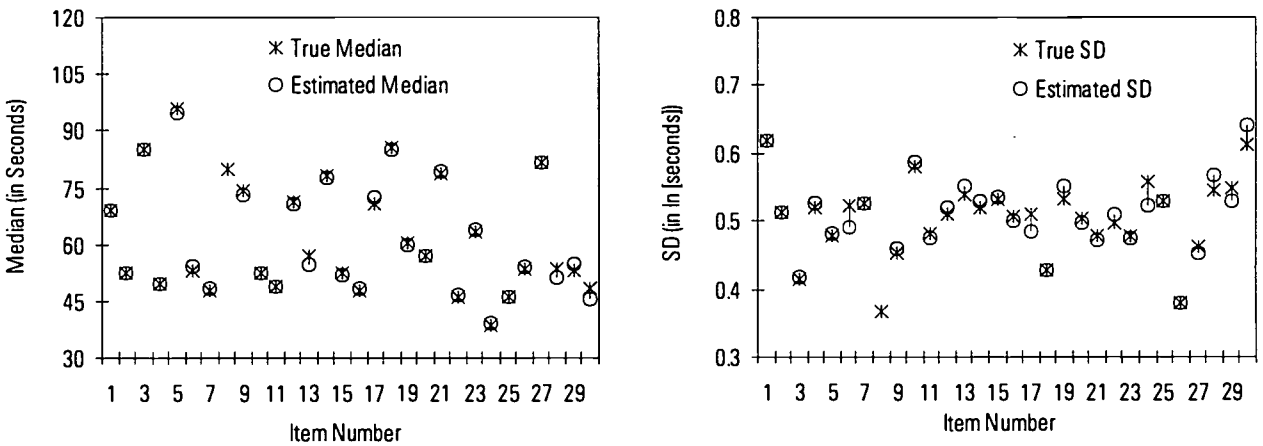

FIGURE 2. *True and estimated median and standard deviation of the solution behavior response time distribution for each item*

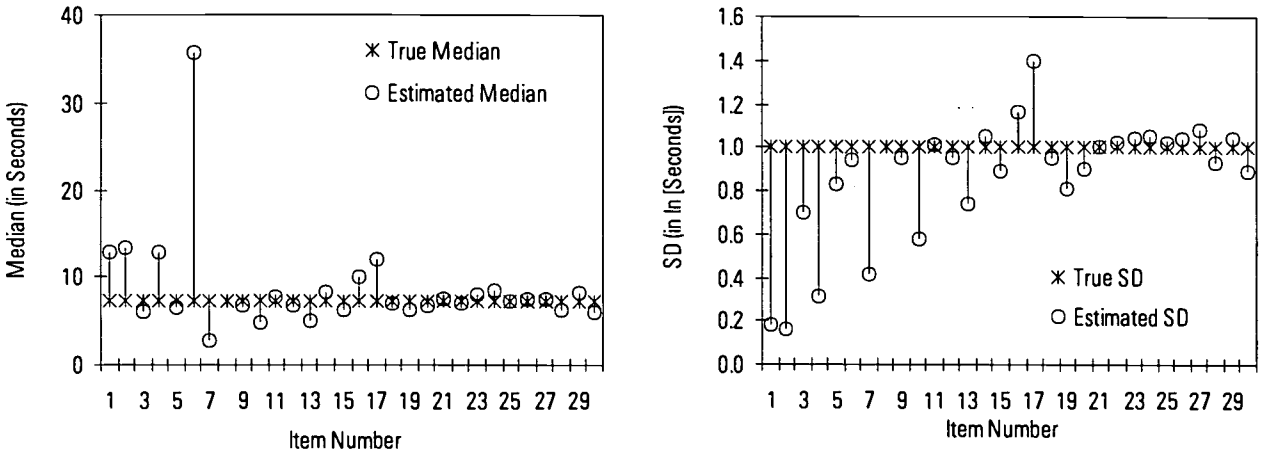

FIGURE 3. *True and estimated median and standard deviation of the rapid-guessing behavior distribution for each item*

Classification Methods

The estimated parameters were used to classify responses as rapid guesses using the following two methods.

Responses classified probabilistically. The first method of classifying responses as rapid guesses attempted to emulate the response-time distribution that would have arisen if there had been no rapid-guessing behavior. That is, the method attempted to sample a solution behavior distribution from the mixture distribution. Based on the mixture model, the proportion of responses at a given response time (in one second intervals) that should be in each distribution was calculated. The responses at each response time were randomly assigned to the two distributions to match the expected proportions. In this sense, the method was probabilistic; there was some probability that a given response would be assigned to a particular state (rapid guessing or solution behavior). The responses that were actually rapid guesses and the responses that were classified as rapid guesses were not exactly the same, of course, but the proportion of responses assigned to each distribution was correct according to the model. (The random assignment to the two underlying distributions was independent of accuracy, so although the proportion of responses classified followed the model, the accuracy rates were not quite right.) The proportion of responses classified as rapid guesses is shown in Table 2 (as is the actual proportion[3] for comparison).

[3]The actual proportion of rapid guesses in the dataset is not ρ (although they are very similar). ρ was used to generate the proportion of rapid guesses, but because of random variability, the numbers are not exactly the same.

TABLE 2

Actual and classified proportions of rapid guesses in the simulated data, and the cutoff (in seconds) used for the cutoff-based approach

Item	Proportion Classified as Rapid Guesses			
	Actual Proportion	Probabilistic Method	Cutoff Method	Value of Cutoff (in seconds)
1	0.000	0.000	0.000	1
2	0.000	0.000	0.000	1
3	0.000	0.000	0.000	1
4	0.001	0.000	0.000	1
5	0.007	0.006	0.006	16.7
6	0.012	0.077	0.017	14.8
7	0.009	0.005	0.005	6.6
8	0.026	-	-	-
9	0.036	0.032	0.029	18.1
10	0.038	0.024	0.024	9.3
11	0.078	0.085	0.068	14.5
12	0.090	0.086	0.077	18
13	0.108	0.083	0.078	12.7
14	0.121	0.125	0.107	21.1
15	0.154	0.134	0.120	14.7
16	0.149	0.177	0.127	16
17	0.158	0.220	0.158	24.6
18	0.155	0.156	0.146	27.5
19	0.179	0.158	0.150	16.4
20	0.175	0.166	0.153	17.6
21	0.178	0.180	0.163	25
22	0.208	0.198	0.167	15.3
23	0.255	0.262	0.232	23
24	0.272	0.299	0.243	15.2
25	0.268	0.264	0.225	16
26	0.281	0.282	0.256	23
27	0.301	0.314	0.288	30
28	0.315	0.282	0.258	16.6
29	0.393	0.427	0.377	22.3
30	0.427	0.381	0.355	15.6

Note. On item 8, a solution was not found. Schnipke and Scrams (1996) had similar trouble on several items when ρ was very small.

Responses classified based on a cutoff. The second method of classifying responses as rapid guesses was based on a cutoff. The cutoffs were established based on the two-state mixture model. The estimated parameters were used to determine, for each item, where the two underlying distributions crossed, weighted by the proportion of test takers in each distribution (ρ_i or 1 - ρ_i). The point at which the distributions cross is where a response is equally likely to come from either the rapid-guessing or solution behavior distribution. This point was used as the cutoff. Once the cutoff for each item was established, all responses with a response time less than the cutoff were classified as rapid guesses, and all responses with a response time greater than the cutoff were classified as solution behavior responses. The cutoffs that were used are given in Table 2, as are the proportions of responses that were classified as rapid guesses using this method.

Treatment of the Responses Classified as Rapid Guesses

Responses classified as rapid guesses were treated as if the item that gave rise to the so-called rapid guess had never been presented to that test taker. To do this, the response code was changed to the "never presented" code (which is not technically true, but it produces the desired result). In this way, responses that appear to be rapid guesses do not influence parameter estimation. Responses classified as solution behavior were not altered. The recoding of responses classified as rapid guesses was done independently (not additively) for the two methods of classification, of course.

Parameter Estimation

BILOG (Mislevy & Bock, 1990) was used to estimate item parameters for the following four conditions:

- no rapid guessing (no rapid guessing was simulated during data generation),
- with rapid guessing (rapid-guessing behavior was simulated during data generation),
- rapid guessing removed probabilistically (rapid-guessing behavior was simulated during data generation; some responses were classified as rapid guesses using the probabilistic method and were recoded as never presented), and
- rapid guessing removed with a cutoff (rapid-guessing behavior was simulated during data generation; some responses were classified as rapid guesses based on a cutoff and were recoded as never presented).

The third and fourth conditions (where rapid guessing was removed) were modifications of the second condition (which included rapid guessing). That is, the exact same responses were used for each simulated test taker item pair, except for the responses that were classified as rapid guesses. For responses classified as rapid guesses, the response code was changed to indicate that the test taker had not received that item, as discussed previously.

In order to set the scale so that the item-parameter estimates from BILOG could be compared, the final estimated ability distribution was scaled to the standard normal. For the condition that includes rapid guessing, forcing the ability distribution to be standard normal will not reflect the true distribution. Although the original ability distribution was standard normal, the inclusion of rapid-guessing behavior distorts the original distribution. Scaling the final estimated ability distribution to the standard normal in this case is not "correct" in the sense that we know the original was distorted. However, in real test data, we would not necessarily know that speed (and hence rapid-guessing behavior) was a factor influencing test scores, and we would scale the final estimated ability distribution to the standard normal. Thus, for the simulated data which included rapid guesses, the ability estimates were scaled as would be done with real test data, and these results were compared to the case in which there was no rapid guessing and to the cases in which an attempt was made to remove rapid guesses.

The IRT item parameters were estimated for each item under each of the four conditions. It was expected that the "no rapid guessing" condition would recover the true item parameters very well and that the "with rapid guessing" condition would not recover the item parameters as well at the end of the test where the proportion of rapid guessing is the highest. How much of a difference there is between these conditions addresses the first research question, How contaminated by speededness are item-parameter estimates?

The second research question, What can be done about the contamination? is addressed by the last two conditions where some responses were classified as rapid guesses and were then removed. The point was to remove responses likely to be rapid guesses based on response times (ignoring the true classifications used to generate those data and which would not be known in real test data) to see if the true item parameters could be recovered more accurately.

Results

Item Difficulty

Figure 4 shows the true difficulty (*b*) parameters, as well as the 4 estimated b values for each item (1 estimated value for each of the 4 conditions). When rapid guesses were included in parameter estimation, the items toward the end of the test appeared more difficult (higher $\hat{b}$ values) than they really are. When there were no rapid guesses and when responses classified as rapid guesses were removed (using either method), the true difficulty parameters were recovered quite well, as shown in Figure 4.

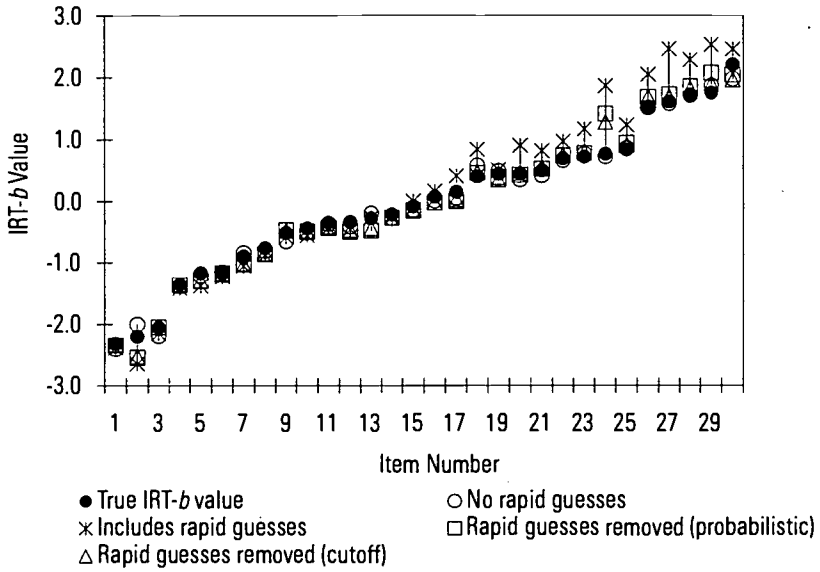

FIGURE 4. *True difficulty (b) parameter and estimates ($\hat{b}$'s) for each condition on each item*

As test-level indices of how well the difficulty parameters were recovered, the correlation between the true and estimated *b*'s and the bias and root mean squared error (*RMSE*) were calculated. As shown in the first row of Table 3, the correlation between *b* and $\hat{b}$ was lowest when rapid guesses were included, highest when there were no rapid guesses, and in between when rapid guesses were removed. All of the correlations were quite high, though, for the difficulty (*b*) parameter.

TABLE 3

Correlations between the true and estimated IRT parameters for the four conditions

			Rapid Guessing Removed	
	No Rapid Guessing	With Rapid Guessing	Probabilistic Method	Cutoff Method
Difficulty (*b*)	.9970	.9888	.9922	.9938
Discrimination (*a*)	.9841	.7876	.9208	.9305
Lower asymptote (*c*)	.5398	.3502*	.5692	.6036

*Not significantly different from 0 at $\alpha = .05$.

The bias statistic was computed for each condition to indicate whether difficulty (*b*) was over- or underestimated overall (i.e., across all items). As shown in the first row of Table 4, *b* was overestimated (large positive bias) when rapid guesses were included (confirming Figure 4). When rapid guesses were removed using either method, there was essentially no bias in *b*, as shown in Table 4, again confirming Figure 4. Likewise, *RMSE*, which indicates the overall amount of error in the estimates across items, was highest when rapid guesses were present and was greatly reduced when rapid guesses were removed using either method, as shown in Table 4.

TABLE 4

Bias and RMSE of the estimated IRT parameters for the four conditions

			Rapid Guessing Removed	
	No Rapid Guessing	With Rapid Guessing	Probabilistic Method	Cutoff Method
Bias				
Difficulty (*b*)	-0.0221	0.1769	0.0023	-0.0128
Discrimination (*a*)	-0.0239	-0.2084	-0.1008	-0.0836
Lower asymptote (*c*)	0.0000	-0.0394	-0.0169	-0.0138
RMSE				
Difficulty (*b*)	0.0085	0.1540	0.0306	0.0221
Discrimination (*a*)	0.0042	0.0846	0.0274	0.0224
Lower asymptote (*c*)	0.0009	0.0032	0.0011	0.0010

Item Discrimination

Figure 5 shows the true and estimated discrimination (*a*) parameters for each item. The true discrimination parameters generally were not recovered as well as the difficulty parameters, as can be seen in the second row of Table 3 (the correlations are lower than those for the *b* parameter). When no rapid guesses were present, the discrimination parameters were recovered fairly well, as indicated by the small values of bias and *RMSE* shown in Table 4. When rapid guesses were included, the discrimination parameters were recovered the least well (as shown by the correlation between a and $\hat{a}$, bias, *RMSE* and visually in Figure 5). The discrimination parameter was underestimated on most items when rapid guesses were present (there was a large negative bias as shown in Table 4). When the responses classified as rapid guesses were removed (using either method), the discrimination parameters were recovered better than when rapid guesses were included, but not as well as the "no rapid guessing" condition, as indicated by bias and *RMSE* and visually (Figure 5).

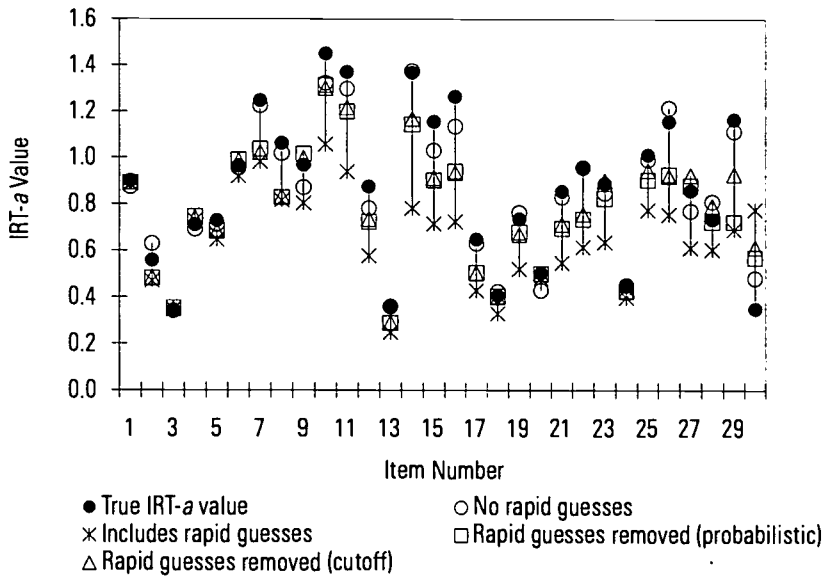

FIGURE 5. *True discrimination (a) parameter and estimates ($\hat{a}$'s) for each condition on each item*

Lower Asymptote

Figure 6 shows the true and estimated lower asymptote (c) parameters for each item. The true lower asymptote parameters were not recovered as well as the difficulty or discrimination parameters (e.g., the correlations between c and $\hat{c}$, shown in the last row of Table 3, are lower than those for difficulty or discrimination). As with the difficulty and discrimination parameters, however, when rapid guesses were included, the lower asymptote parameters were recovered the least well (the correlation between c and $\hat{c}$ was the lowest as shown in Table 3, and bias and *RMSE* were the highest as shown in Table 4). When the responses classified as rapid guesses were removed (using either method), the lower asymptote parameters were recovered at least as well as when no rapid guesses were present, as shown in Figure 6 and Tables 3 and 4.

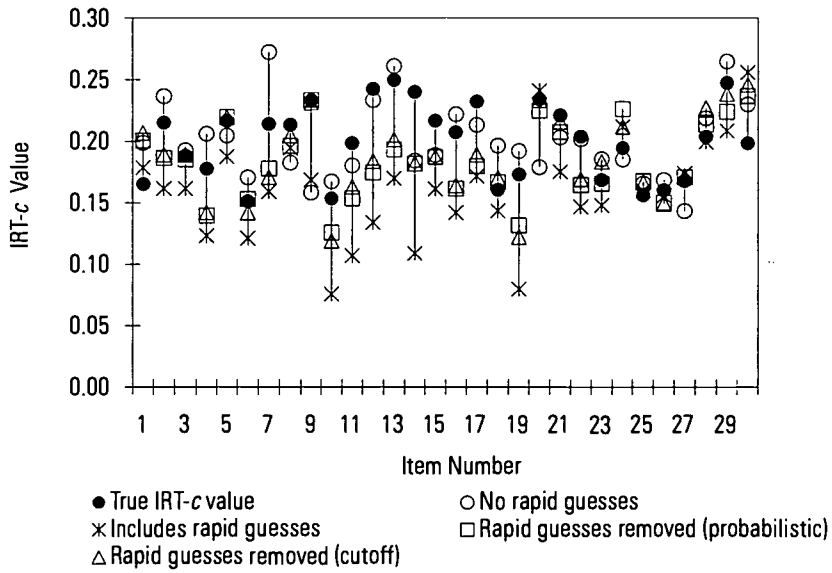

FIGURE 6. *True lower asymptote (c) parameter and estimates ($\hat{c}$'s) for each condition on each item*

Item Characteristic Curves

To compare the combined effects of the IRT *a*, *b*, and *c* parameters, the ICC for each item was calculated. The ICC provides the probability that a test taker with a given ability will answer the item correctly. Figure 7 shows the true and estimated ICCs for several items throughout the simulated test (items 10, 15, 20, 25, and 30). As shown in Figure 7, when rapid guesses were included, the ICC was artificially low for items toward the end of the test; a test taker appears to have a smaller probability of answering an item correctly than is really the case. This, of course, corresponds to the difficulty parameter being too high. (The slopes of the ICCs were also altered by rapid-guessing behavior, which is related to the *a*'s and *c*'s being recovered less well.) When there were no rapid guesses, or when responses classified as rapid guesses were removed (using either method), the ICCs were recovered very well.

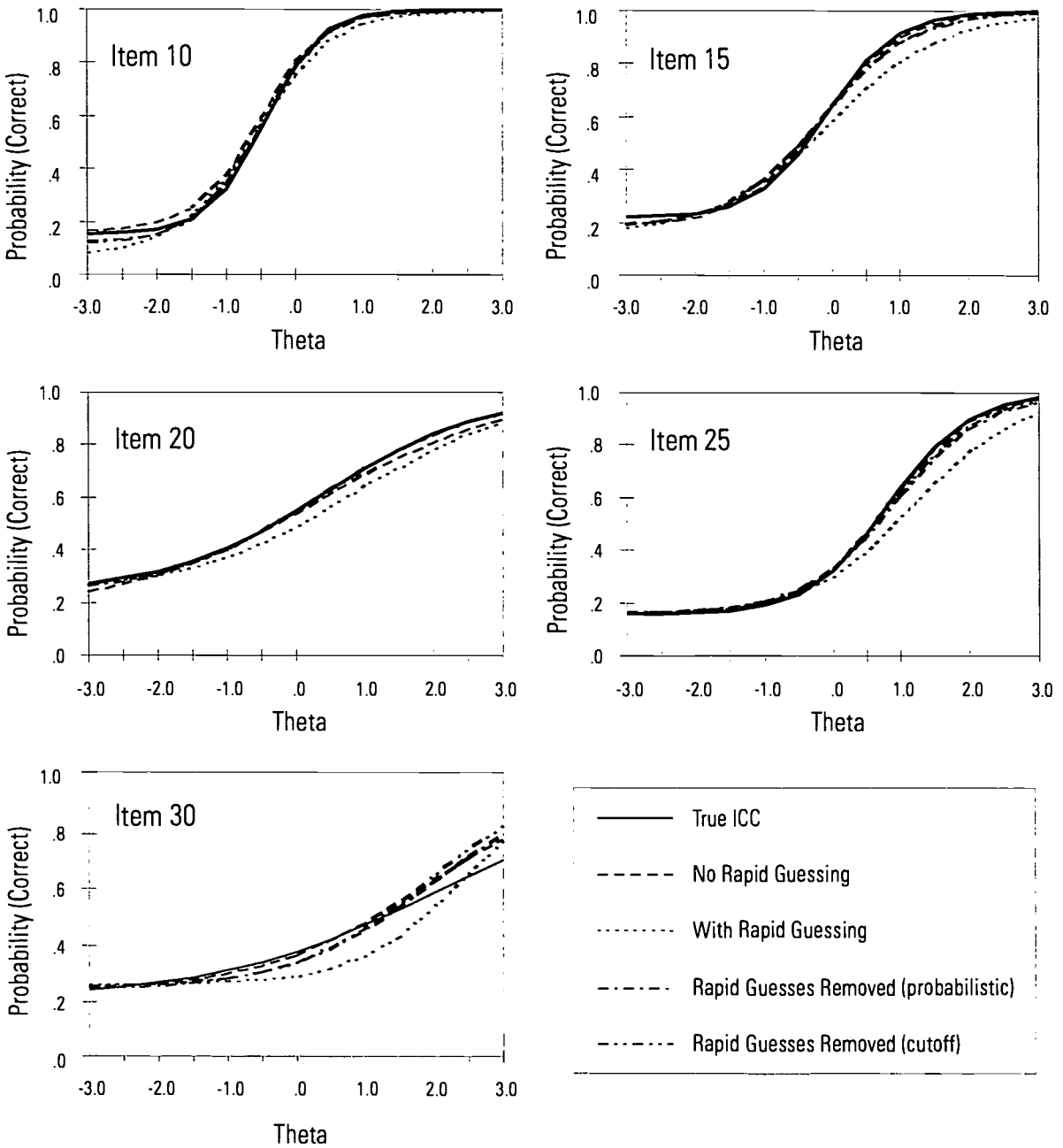

FIGURE 7. *Estimated and true item characteristic curves for several simulated items*

Item Information Functions

Because item information is often used to select items for both linear and adaptive tests, it is useful to determine the effects of rapid-guessing behavior on item information functions. Figure 8 shows the information functions for the same items shown in Figure 7. On most items, when rapid-guessing behavior was present, the information functions were underestimated; the items appear to be less informative than they really are. This is related to the discrimination (*a*) parameters being underestimated. On item 30, the *a* parameter was overestimated when rapid-guessing behavior was present (as shown in Figure 5), and item information was correspondingly inflated. On all items toward the end of the test, the ability (θ) level at which the item provides maximum information was inflated when rapid-guessing behavior was present (because the items seemed more difficult). When rapid-guessing behavior was removed using either method, item information (including the θ level at which the item provides maximum information) was recovered more accurately for all items.

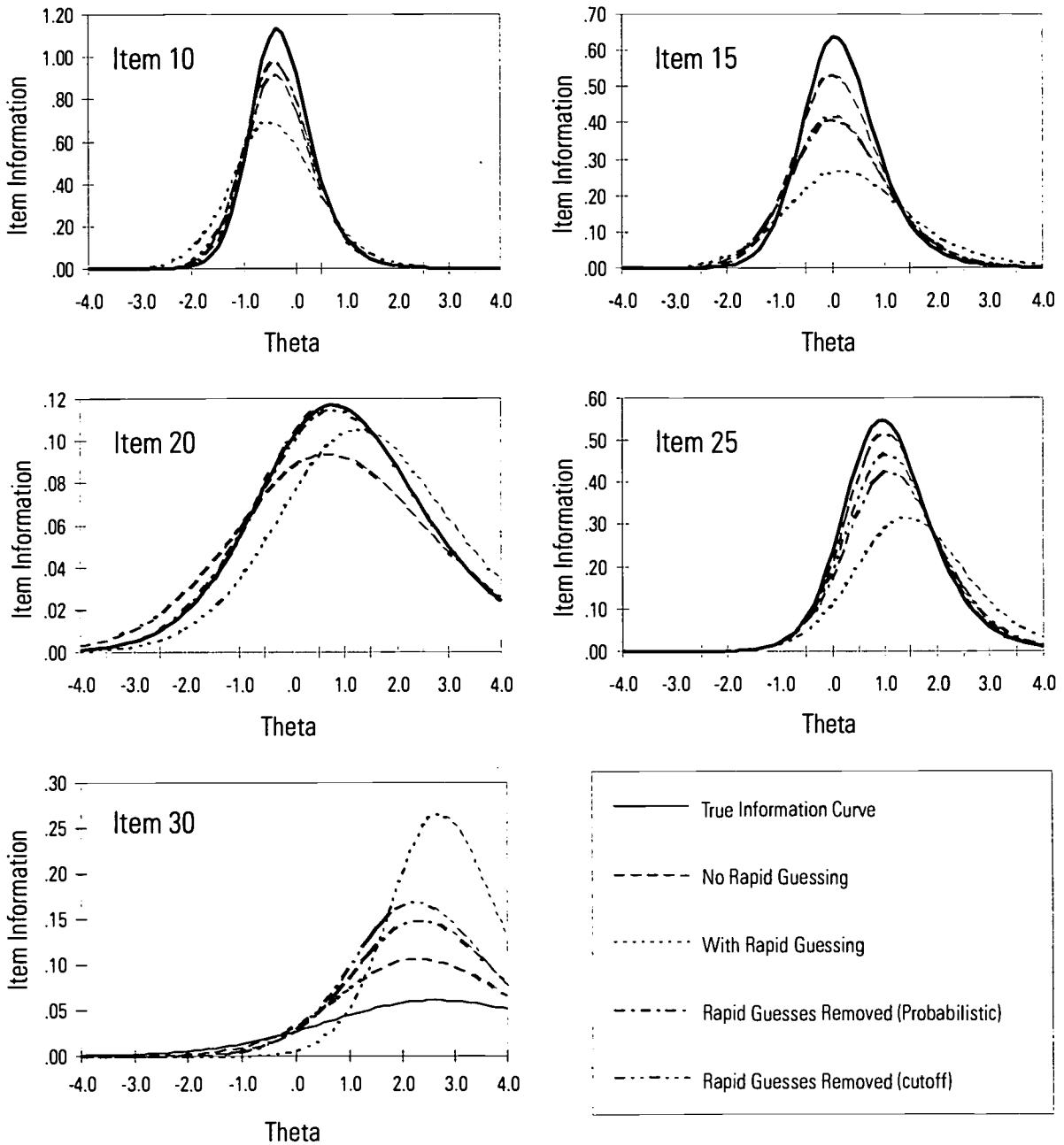

FIGURE 8. *Estimated and true item information functions for several simulated items*

Discussion

On speeded multiple-choice tests such as the LSAT, some test takers are likely to respond very quickly to items as time expires in the hopes of getting some items right by chance. This is often a wise strategy for test takers to use, but these rapid responses cause problems for practitioners who analyze test data. These fast responses, or "rapid guesses," will distort item-parameter estimates if they are included during parameter estimation. As shown in the present study, when rapid guesses are present, items appear more difficult and less discriminating than they really are.

When tests are administered on computers, response times can be collected. Response times provide clues about which responses are rapid guesses and which are not. In the present study, response times were simulated to match distributions found by Schnipke and Scrams (1997) and were used to classify responses as rapid guesses to see if removing these responses would provide more accurate parameter estimates. A simulation was required because the true parameters needed to be known.

Two methods were used to classify responses as rapid guesses. One attempted to emulate a solution-behavior-only distribution by sampling from the mixture distribution. Some, but not all, of the fastest response times were removed so that the remaining number of responses at each response time matched the number expected in the underlying solution-behavior distribution. The other method used a cutoff to classify responses. The cutoff was placed where a response was equally likely to come from either underlying distribution. Responses made faster than the cutoff were more likely to come from the rapid-guessing distribution and therefore were classified as rapid guesses. This method truncated the distribution, whereas the first method did not.

The two methods for classifying responses as rapid guesses produced virtually identical results. The parameter estimates were recovered equally well with the two methods. Because the cutoff approach is easier to implement and easier to understand, the cutoff approach is recommended, although the other approach works well, too.

When new tests are constructed, the new tests may not have the desired psychometric properties if inaccurate item parameters (e.g., ones derived for items at the end of a speeded test) are used. This may be a problem especially in adaptive tests because item selection and test taker ability estimation depend heavily on item parameters. If response times are available, it is recommended that practitioners look for evidence of rapid-guessing behavior in their data and correct for it if it is there, especially if the items will be used in adaptive tests or if the item will change position in subsequent forms of the test.

There are several limitations to the present study; these include the following: (1) ability and rapid-guessing behavior were generated independently; (2) solution-behavior response-time distributions were generated independently of item difficulty; (3) the test design was a nonadaptive fixed-length test; and (4) the proportion of rapid guesses and item difficulty both increased as a function of item position. For generalizing the results, the biggest limitation is probably that ability and rapid-guessing behavior were simulated independently. Ability and rapid-guessing behavior probably are related in real data, but because the exact nature of this relationship is not known, no dependency was incorporated. Future work will investigate the relationship between ability and rapid-guessing behavior and between solution-behavior response-time distributions and item difficulty.

In terms of test design, a nonadaptive fixed-length test was used because that is the simplest case, and as such, provided a good starting place. Future work will consider an adaptive design. There are test designs other than the nonadaptive fixed-length design that would provide data that could be analyzed with the techniques used here. For instance, a design that uses testlets may provide appropriate data. A testlet refers to a group of items that are administered as a unit (Wainer & Kiely, 1987). If the items within the testlets are not adaptively administered, items toward the end of each testlet may contain rapid-guessing behavior. One can imagine rearranging items into new testlets based on item parameters, and we would want the parameter estimates to be free of rapid-guessing behavior. Designs based on such testlets include Lewis and Sheehan's (1990) Bayesian mastery testing design; Luecht, Nungester, and Hadadi's (1996) computer administered sequential testing design; and Reese and Schnipke's (1996) two-stage testlet design.

Oshima (1994) found random guessing distorts difficulty estimates the least when items are arranged in order of ascending difficulty (if random guessing also increases toward the end of the test). This was the case in the present study (both item difficulty and rapid guessing increased toward the end of the test). The difficulty estimates would be expected to be more distorted by rapid guesses if items were not increasingly difficult toward the end of the test, and in this situation it might be even more important to counteract the effects of rapid-guessing behavior by attempting to remove such responses.

References

Evans, M., Hastings, N., & Peacock, B. (1993). *Statistical distributions*. New York: John Wiley & Sons.

Lewis, C., & Sheehan, K. (1990). *Using Bayesian decision theory to design a computerized mastery test* (Research Report No. RR-90-28). Princeton, NJ: Educational Testing Service.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. New York: Oxford University.

Luecht, R. M., Nungester, R. J., & Hadadi, A. (1996, April). *Heuristic-based CAT: Balancing item information, content and exposure*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Mislevy, R. J., & Bock, R. D. (1990). BILOG 3: Item analysis and test scoring with binary logistic models [Computer software and manual]. Chicago: Scientific Software, Inc.

Norušis, M. J. (1994). *SPSS advanced statistics™ 6.1*. Chicago: SPSS, Inc.

Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement, 31*, 200-219.

Reese, L. M., & Schnipke, D. L. (1996, June). *An evaluation of a two-stage testlet design for CAT*. Paper presented at the annual meeting of The Psychometric Society, Banff, Alberta, Canada.

Schnipke, D. L. (1995, April). *Assessing speededness in computer-based tests using item response times*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement, 34*, 213-232.

Yamamoto, K. (1995). *Estimating the effects of test length and test time on parameter estimation using the HYBRID model*. (TOEFL Technical Report No. TR-10). Princeton, NJ: Educational Testing Service.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185-201.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

[X] This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

[ ] This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").