# DOCUMENT RESUME

ED 467 806 TM 034 347

| | |
|---|---|
| AUTHOR | Schnipke, Deborah L.; Scrams, David J. |
| TITLE | Modeling Item Response Times with a Two-State Mixture Model: A New Approach to Measuring Speededness. Law School Admission Council Computerized Testing Report. LSAC Research Report Series. |
| INSTITUTION | Law School Admission Council, Princeton, NJ. |
| REPORT NO | LSAC-R-96-02 |
| PUB DATE | 1999-03-00 |
| NOTE | 22p. |
| PUB TYPE | Reports - Research (143) |
| EDRS PRICE | EDRS Price MF01/PC01 Plus Postage. |
| DESCRIPTORS | Admission (School); *College Entrance Examinations; *Guessing (Tests); *Law Schools; *Timed Tests |
| IDENTIFIERS | Graduate Record Examinations; *Law School Admission Test; Mixtures; *Speededness (Tests) |

ABSTRACT

Speededness refers to the extent to which time limits affect test takers' performance. With regard to the Law School Admission Test (LSAT), speededness is currently measured by calculating the proportion of test takers who do not reach each item on the test. These proportions typically increase slightly toward the end of the test, indicating that the LSAT is partially speeded. Because the LSAT is number-right scored (i.e., no points are subtracted for incorrect responses), test takers are encouraged to guess on items rather than leave them blank. Therefore, this measure of speededness for the LSAT probably underestimates the true amount of speededness on the test. A more accurate assessment of speededness should also reflect the tendency of test takers to guess rapidly on items as time expires. This "rapid guessing" component of speededness can be estimated by modeling response-times with a two-state mixture model, as demonstrated with data from a computer-administered reasoning test, the analytical measure of the Graduate Record Examination for 7,218 test takers. The combined effect of unreached items and rapid guessing provides a more complete measure of speededness than has previously been available. (Contains 1 table, 8 figures, and 17 references.) (Author/SLD)

LSAC RESEARCH REPORT SERIES

TM

# Modeling Item Response Times With a Two-State Mixture Model: A New Approach to Measuring Speededness

Deborah L. Schnipke and David J. Scrams
Law School Admission Council

Law School Admission Council
Computerized Testing Report 96-02
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

Speededness refers to the extent to which time limits affect test takers' performance. Speededness has traditionally been defined as the extent to which some test takers are unable to finish all the test items, and it has been measured by the percentage of test takers who do not reach a certain number of items (e.g., all of the items, or 75% of the items). With regard to the Law School Admission Test (LSAT), speededness is currently measured by calculating the proportion of test takers who do not reach each item on the test. These proportions typically increase slightly toward the end of the test, indicating that the LSAT is partially speeded. The definition ignores the fact that random guessing is likely to occur as time expires, especially on multiple-choice tests, such as the LSAT, that do not subtract points for wrong responses. In fact, coaching schools, and often the test instructions, encourage test takers to rapidly fill in any remaining answers on the answer sheet, rather than leave them blank. Therefore, our measure of speededness probably underestimates the true amount of speededness on the test. The primary function of the LSAT is not to measure rate of work (speed), thus speededness is considered an ancillary variable. To know how great an effect this ancillary variable has on test scores, an accurate measure of speededness is needed.

Test takers are not likely to rapidly fill in any remaining answers on the answer sheet if they have enough time to read and fully consider each item. Thus, the presence of "rapid-guessing behavior" indicates that the test was speeded for the test taker. Traditional indices underestimate the extent of speeded behavior because test takers who rapidly guess do not have "unreached" items. To accurately measure speededness, rapid guesses must be reflected in the estimate.

If test takers had enough time, they would presumably engage in "solution behavior" on all items. In solution behavior, test takers actively try to determine the correct answer (solution) to every item; they read each item carefully and fully consider the answer. Response-times arising from solution behavior will likely depend on item length, difficulty, and other characteristics, as well as person-specific variables.

In contrast, when engaging in rapid-guessing behavior, test takers may skim items briefly for key words, but they do not completely read items. Consequently, item characteristics, such as difficulty, length, and content, may have little effect on response-times arising from rapid-guessing behavior.

Clearly, rapid-guessing behavior and solution behavior will lead to different response-time distributions. In the present study, a rigorous measure of speededness is developed by mathematically modeling item response-time distributions from a computer-administered version of the Graduate Record Examinations (GRE) General Test. The data analyzed are from the analytical measure of the GRE which is comprised of two item types (logical reasoning and analytical reasoning). These two item types are very similar to two of the item types (with the same name) on the current LSAT. Using standard model-fitting techniques, it was determined that item response-time distributions on the first half of the test section could be described by a single underlying distribution, whereas items on the last half of the test required two underlying distributions. Additional evidence suggests that on the first half of the test section, the response-times were the result of solution behavior, and on the last half of the test section the two underlying distributions were the result of solution behavior and rapid-guessing behavior. The modeling techniques were successful in detecting items containing rapid-guessing behavior and in estimating the proportion of responses that were rapid guesses. Combining the proportion of rapid guesses on each item with the proportion of test takers who did not reach each item provides a new, rigorous measure of speededness which can be used to better understand what test scores mean.

The Law School Admission Council (LSAC) is currently researching the feasibility and advisability of administering the LSAT on computer. If the LSAT is transferred to a computer-administered format, the methodology developed in this paper will be useful for providing an accurate measure of speededness on the LSAT.

## Abstract

Speededness refers to the extent to which time limits affect test takers' performance. With regard to the Law School Admission Test (LSAT), speededness is currently measured by calculating the proportion of test takers who do not reach each item on the test. These proportions typically increase slightly toward the end of the test, indicating that the LSAT is partially speeded. Because the LSAT is number-right scored (i.e., no points are subtracted for incorrect responses), test takers are encouraged to guess on items rather than leave them blank. Therefore, this measure of speededness for the LSAT probably underestimates the true amount of speededness on the test. A more accurate assessment of speededness should also reflect the tendency of test takers to rapidly guess on items as time expires. This "rapid-guessing" component of speededness can be estimated by modeling response-times with a two-state mixture model, as demonstrated with data from a computer-administered reasoning test. The combined effect of unreached items and rapid guessing provides a more complete measure of speededness than has previously been available.

## Introduction

Testing theorists distinguish between tests that measure "power" and tests that measure "speed." Items in a pure power test range in difficulty, and there is no time limit; the goal is to measure how accurately test takers can answer the items. The Information subtest of the Wechsler Adult Intelligence Scale (WAIS), which tests for general knowledge, is a good example of a pure power test. In contrast, items in a pure speed test are very easy, and the time limit is strict; the goal is to measure how quickly test takers can respond. An example of a speed test is the Digit Symbol subtest of the WAIS in which test takers match symbols to numbers as quickly as possible (accuracy does count as well). Most aptitude and assessment tests are designed essentially as power tests, but a speed component is introduced when these tests are administered with a time limit—a practice often adopted for group administration. The Law School Admission Test (LSAT) is such a test.

Speededness, or the extent to which a test is speeded, is of concern primarily because it can adversely affect test characteristics (e.g., reliability; Gulliksen, 1950) and item parameters (Mollenkopf, 1950; Oshima, 1994; Schnipke, 1996). On a more theoretical note, Gulliksen (1950) pointed out that the item indices of classical test theory were developed for power tests and might not be appropriate if the speed component is too large. Similarly, Hambleton and Swaminathan (1985) argued that unidimensional item response theory (IRT) models implicitly assume that the test is unspeeded; speed and power components would require separate dimensions.

The Law School Admission Council (LSAC) is currently researching the feasibility and advisability of administering the LSAT on computer. One advantage of computerized tests is that response-times are available operationally. The primary function of the LSAT is not to measure rate of work (speed), thus speededness is considered an ancillary variable. To know how great an effect this ancillary variable has on test scores, an accurate measure of speededness is needed. In this study, a rigorous measure of speededness is developed using item response-time data from a computer-administered version of the Graduate Record Examinations (GRE) General Test. The data analyzed are from the analytical measure of the GRE, which is comprised of two item types (logical reasoning and analytical reasoning). These two item types are very similar to two of the item types (with the same name) on the current LSAT. If the LSAT is transferred to a computer-administered format, the methodology developed in this paper will be useful for providing accurate measures of speededness.

The data for this project were provided to the first author by the Educational Testing Service (ETS) and the Graduate Record Examinations (GRE) Board through the GRE Graduate Research Assistantship in Psychometrics Program. Support for the second author was provided by a National Science Foundation predoctoral fellowship.

### *Measuring Speededness*

Theorists have assumed that on a pure power test (one that is completely unspeeded), all test takers will answer all test items. Consistent with this assumption, some researchers have classified a test as speeded if some test takers did not reach[1] all items; this is a valid conclusion based on the logic of *modus tollens*. Unfortunately, researchers have also classified tests as pure power as long as all test takers completed all items; this is an example of *Affirming the Consequent*, a logical fallacy. It may be possible for test takers to be adversely affected by the time limit but to still give a response to every item.

Other researchers switched from dichotomous decisions (power vs. speeded) to indices of speededness that are based on the percentage of test takers who do not reach all items.[2] This is the current measure of speededness employed by LSAC with regard to the LSAT. These indices are plagued by a more complex version of *Affirming the Consequent*. Namely, the indices are based on the assumption that the extent to which test takers fail to complete all the items adequately represents the extent of speededness. Repercussions of time pressure that do not result in unreached items will not be reflected in these indices.

In particular, this treatment of speededness ignores the fact that random guessing is likely to occur as time expires, especially on multiple-choice tests, such as the LSAT, that do not penalize wrong responses. In fact, coaching schools, and often the test instructions, encourage test takers to rapidly fill in remaining answers on answer sheets rather than leave them blank. Test takers are not likely to engage in this behavior if they have enough time to read and fully consider each item, so the presence of "rapid-guessing behavior" indicates that the test was speeded for the test taker. Traditional indices underestimate the extent of speeded behavior because test takers who rapidly guess do not have "unreached" items.

Yamamoto (1995) was also concerned with the guessing component of speededness. His primary goal was to develop a method of estimating item parameters that eliminates the effects of random responding. He assumed that test takers begin a test in what we will call "solution behavior." In solution behavior, test takers actively try to determine the correct answer (solution) to every item. Yamamoto assumed that test takers might switch to a random-responding strategy (what we call "rapid-guessing behavior") as time elapses, and he identified such behavior by the lowered accuracy of the responses. The current work is similar to Yamamoto's, but we focus on response times, rather than accuracy, to categorize test taker behavior.

### *Current Model*

Like Yamamoto (1995), we assume that test takers choose to engage in either solution behavior or rapid-guessing behavior. Further, we assume that test takers can switch strategies at any point and that they do so in response to the time constraints on the test. When engaged in solution behavior, test takers read each item carefully and fully consider their answers. Accuracy will then depend jointly on test taker ability and item difficulty and other item characteristics. Consequently, response times arising from solution behavior will likely depend on item length, difficulty, and other characteristics, as well as person-specific variables.

In contrast, when test takers respond rapidly to items as time expires, their accuracy rates will be at or near chance because they are not fully considering the items. In rapid-guessing behavior, test takers may skim items briefly for key words, but they do not thoroughly read the items. Consequently, item characteristics, such as difficulty, length, and content, may have little effect on response times arising from rapid-guessing behavior.

[1] Unreached items on a paper-and-pencil test are defined as unanswered items at the *end* of a timed test section. Unanswered items that are followed by at least one answered item are defined as omitted items. Notice that on a computer-administered test, more precise definitions are possible. Omitted items are items that are presented to the test taker (i.e., appear on the screen) but are unanswered. Unreached items are items that were never presented to the test taker.

[2] There are other related indices, such as the percentage of test takers who do not reach 75% of the items and the percentage of items that were not reached by 100% of the test takers (Swineford, 1956) and the ratio of the variance of unreached items to the total error variance (Gulliksen, 1950). For simplicity, indices such as these will not be discussed, but the same logic applies.

The above processing assumptions lead to the following predictions: First, whereas the response-time distributions for solution behavior may vary substantially across items, the response-time distributions for rapid-guessing behavior will be similar across items. Second, the tendency of test takers to engage in rapid-guessing behavior will increase as time expires. Third, accuracy rate as a function of response-time will be monotonically increasing over the range of response-times common to both distributions. These hypotheses will be elaborated on in a subsequent section.

## *Mathematical Development*

### *Two-State Mixture Model*

If each test taker is assumed to be engaged in either rapid-guessing or solution behavior on a given item, and these two types of behavior are assumed to be associated with different response-time distributions, then the observed response-time distribution can be described as a mixture of rapid-guessing and solution-behavior response-time distributions. This is a two-state mixture model (e.g., Luce, 1986; Townsend & Ashby, 1983) which can be expressed mathematically as

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

where

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

$\rho_i$ 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.

Parameters of the model can be estimated from data as long as the shapes of the underlying distributions can be specified reasonably and $\rho_i$ is not too close to 0 or 1. As $\rho_i$ approaches 0, the parameters of the rapid-guessing distribution cannot be estimated because there are too few rapid-guessing observations. Similarly, as $\rho_i$ approaches 1, the parameters of the solution-behavior distribution cannot be estimated.

### *Single-State Model*

When $\rho_i = 0$, the standard mixture model simplifies to

$$F_{Oi} = F_{Si} \tag{2}$$

because all test takers are engaged in solution behavior. This would be expected for items presented at the beginning of a speeded test and all items on an unspeeded (power) test. When $\rho_i = 1$, all test takers are engaged in rapid-guessing behavior (an unlikely finding). More generally, $\rho_i$ can be used as an index of the extent of rapid-guessing behavior (i.e., one component of an item-level index of speededness).

On an unspeeded test, the single-state model given in Equation 2 should account for the full distribution of response times. If some test takers are engaging in rapid-guessing behavior, this model will be unable to account for the number of fast responses (i.e., the rapid guesses). The two-state standard mixture model, given in Equation 1, will be needed to account for the response-time distribution on items with rapid guessing behavior.

## *Hypotheses*

### *Common-Guessing Distribution*

As noted above, item characteristics should not affect response times in the rapid-guessing distribution. Therefore, the rapid-guessing distribution should be the same for all items that have rapid-guessing behavior. To test this hypothesis, a "common-guessing mixture" model was constructed by constraining the parameters of the rapid-guessing distribution in the standard mixture model to be constant across items. The proportion of rapid guesses ( $\rho_i$ ) and the solution-behavior parameters were allowed to vary across items. The mathematical form of the common-guessing mixture model is given by

$$F_{Oi} = \rho_i F_{CG} + (1-\rho_i) F_{Si} \tag{3}$$

where $F_{CG}$ is the common-guessing mixture distribution and is the same for all items.

The only difference between the common-guessing mixture model and the standard mixture model (Equation 1) is that $F_{CG}$ does not vary across items, whereas $F_{Gi}$ does vary across items. Thus, the common-guessing mixture model is a special case of the standard mixture model. (Note that the single-state model in Equation 2 is also a special case of the standard mixture model where $\rho_i = 0$ for all items.) If the common-guessing mixture model fits the data as well as the standard mixture model, the hypothesis that rapid-guessing behavior is the same across items is supported.

### *Increasing $\rho_i$ by Position Function*

Because rapid-guessing behavior is a consequence of time constraints, the proportion of test takers engaging in rapid-guessing behavior should increase as time expires. This assumption, and the assumption that test takers answer items in order, lead to the prediction that $\rho_i$ will be an increasing function of item position. Examination of the best-fitting parameter estimates (i.e., the $\hat{\rho}_i$ 's) should provide evidence bearing on this hypothesis.

### *Increasing Accuracy by Response-Time Function*

Rapid-guessing behavior should result in lower accuracy and faster response times than does solution behavior. The response-time distributions associated with rapid-guessing and solution behavior are assumed to be typical response-time distributions; that is, unimodal distributions with central mass and decreasing tails (Luce, 1986). If these distributions overlap, accuracy should be a monotonically increasing function of response time across the range of overlap because the proportion of responses arising from solution behavior is also an increasing function of response time. (If we were willing to assume that accuracy and response time are independent when conditioned on behavior, or if we could specify the nature of the relationship, we could offer a more specific prediction, but such theorizing is probably premature.) Examination of the accuracy by response-time functions will provide data for the present hypothesis as well as future theoretical work.

To determine whether the mixture-model approach is a reasonable way to model speededness and to test the three hypotheses about rapid-guessing behavior, the single-state, standard mixture, and common-guessing mixture models were fit to the response-time distribution for each item on a computer-administered version of the Graduate Record Examinations (GRE) General Test. The single-state model should fit the response times for an item if there is no rapid-guessing behavior, and it should underpredict the number of fast responses if there is rapid-guessing behavior. On such items, the standard mixture model or the common-guessing mixture model will be required to describe the entire distribution of response times accurately.

## Method

The computer-administered, nonadaptive GRE General Test analyzed in the present study was administered to 7,218 test takers in the 1992-1993 academic year. The test takers chose to take the computerized version rather than the paper-and-pencil version of the test. Because scores were used operationally (i.e., scores were reported to graduate schools for use in admission decisions), test takers were highly motivated to do well. Because no points were subtracted for incorrect answers, it was in the test takers best interest to guess at answers rather than leave them blank. Test takers were aware that no penalties for incorrect responses would be applied.

The analytical section of the GRE that was analyzed in the present study consisted of 25 items and was administered with a 32-minute time limit. All test takers received the same 25 items. The analytical section contained two item types: logical reasoning items, which measure the ability to understand, analyze, and evaluate arguments (items 7, 8, 9, 23, 24, and 25), and analytical reasoning items, which occur in sets and measure the ability to understand a given structure of arbitrary relationships and to deduce new information from the relationships given (set 1, items 1–6; set 2, items 10–14; set 3, items 15–18; and set 4, items 19–22). Items were presented one at a time on the computer screen. For the analytical reasoning items, the common stimulus (which contained the structure of the arbitrary relationships) was displayed on half the screen, and the items that referred to it were displayed one at a time on the other half of the screen.

Although, by default, the items were presented in numerical order, test takers could answer the items in any order by using "navigation tools." They could omit items (i.e., see an item but not answer it and proceed to another item), and they could go back to previously viewed items and change their answers. The test was administered in this way so that it would be as similar as possible to the paper-and-pencil version.

### *Response-Time Collection*

Because test takers could "navigate" through the test in any order, they could attempt an item more than once. The recorded response time for an item was the total time spent on the item during all attempts to that item. The recorded response was the answer provided on the final attempt on the item. (Multiple attempts were uncommon; the mean number of attempts on each item ranged from 1.12 attempts on item 1, to 1.54 attempts on item 25, excluding those test takers who never saw/attempted[3] the item.) Response times were collected to one-second precision.

## Results

Probability density functions (PDFs) for the response times to several representative items[4] are shown in Figure 1. Response time is plotted on the abscissa, and the proportion of test takers who responded at each level of response time is plotted on the ordinate. The graphs are stacked area charts (at each response-time level, correct responses are stacked on top of incorrect responses, not behind them).

All the response-time distributions are positively skewed, and many are unimodal (items on the first half of the test[5] tended to be unimodal; see item 4 in Figure 1 for an example). Positive skewness and unimodality are characteristic of response-time distributions in general, but some of the distributions are uncharacteristically bimodal due to an unusually large number of fast response-times. This is true of all items on the last half of the test (items 13–25; see Figure 1 for examples). Bimodality is characteristic of some mixture distributions (but see Townsend & Ashby, 1983, for exceptions). The fast responses also tend to be incorrect, consistent with the claim that they arise from rapid-guessing behavior.

[3] The item never appeared on the screen in this case.

[4] Although only a limited number of items can be shown in the figures, all of the items are described in general terms.

[5] For convenience, we refer to the analytical section from the GRE as "the test," although it was one of seven components, or sections, of the entire GRE.

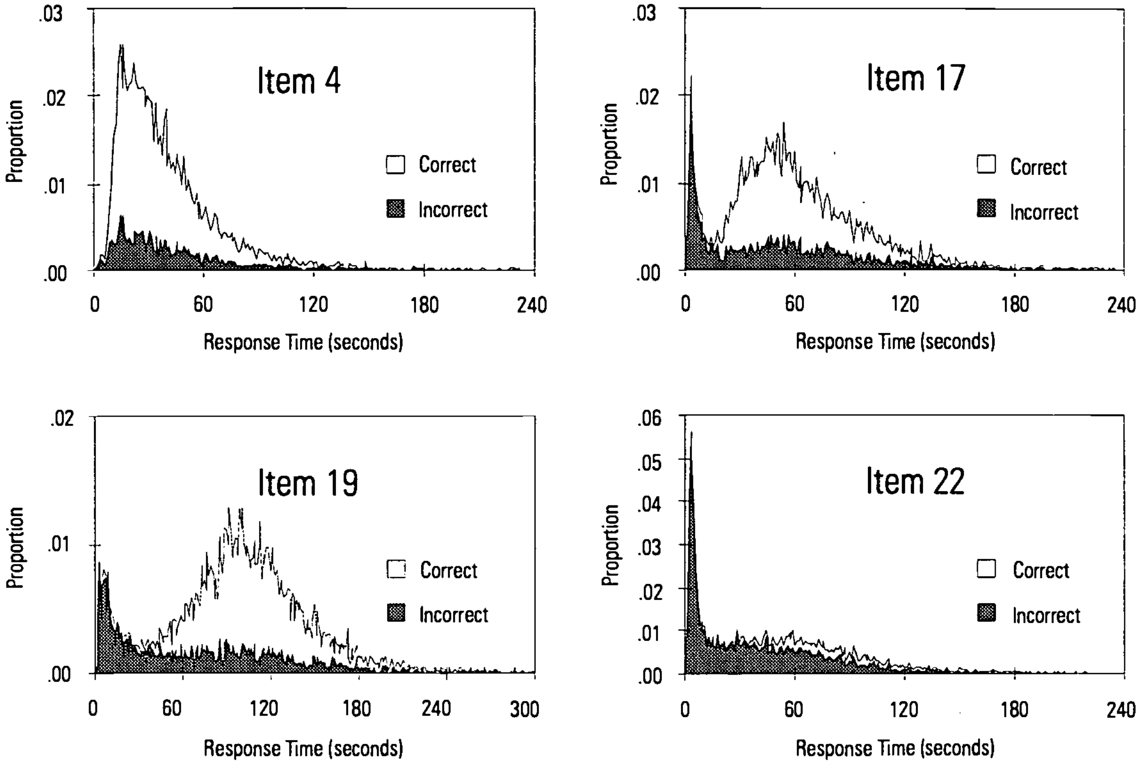

FIGURE 1. *Probability density functions of response times for several representative items. The scales are not the same across items.*

*Modeling Response Times*

Three models were fit to the observed cumulative distribution function (CDF)[6] for each item: (1) a single-state model (Equation 2), which would be expected to fit typical response-time data; (2) a two-state standard mixture model (Equation 1), which will fit much better than the single-state model if there are two underlying distributions (e.g., a solution-behavior distribution and a rapid-guessing distribution); and (3) a two-state standard mixture model that is constrained to have a common rapid-guessing distribution for all items (Equation 3). The models were fit to the observed CDF for each item using nonlinear regression. The Levenberg-Marquardt algorithm is used by SPSS for unconstrained models (Norušis, 1994) and was used for the single-state and standard mixture models. The sequential quadratic programming algorithm is used by SPSS when constraints are placed on the model; it was used for the common-guessing mixture model where the $\rho_i$'s were constrained to be non-negative.

*Choosing the Shapes of the Underlying Distributions*

In order to fit the models to data, the shapes of the solution and rapid-guessing response-time distributions must be specified. We assumed that both distributions would include the unimodality and positive skewness typical of response-time distributions in general (Luce, 1986). In the absence of theoretical predictions, we assumed that the distributions would be approximated well by lognormal distributions. This distribution was selected because the parameters are fairly intuitive, and thus easier to interpret than parameters of some other distributions (e.g., the Ex-Gaussian). The lognormal probability density function is given by

$$f(t) = \frac{1}{\sqrt{t\sigma(2\pi)}} \text{ x } \exp\left\{ \frac{-[ln(t/m)]^2}{2\cdot\sigma^2} \right\} \qquad (4)$$

[6] The cumulative distribution function (CDF) was used because it tends to be a more stable estimator than the probability density function (PDF). Equations 1, 2, and 3 hold for both the CDF and the PDF.

where

$t$ is the response time (RT),
$m$ is the scale parameter, the median of the RTs, and
$\sigma$ is the shape parameter, the standard deviation of the ln(RTs) (Evans, Hastings, & Peacock, 1993).

The median, $m$, is expressed in seconds, and $\sigma$ is expressed in ln(sec).

By fitting $F_{Gi}$, $F_{Si}$, and $F_{CG}$ (from Equations 1, 2, and 3) as lognormal distributions, each underlying distribution could be specified with only two parameters. This results in two parameters per item for the single-state model ($m$ and $\sigma$) for a total of 50 parameters (2 parameters x 25 items). There were five parameters per item for the standard mixture model (two $m$'s, two $\sigma$'s, and $\rho$) for a total of 125 parameters (5 parameters x 25 items). Finally, there were three parameters per item for the common-guessing mixture model ($m_S$, $\sigma_S$, and $\rho$, where the $S$ subscripts refer to the solution-behavior distribution), in addition to $m_G$ and $\sigma_G$ (where the $G$ subscripts refer to the rapid-guessing distribution) from the common-guessing mixture distribution, for a total of 77 parameters (3 parameters x 25 items, plus 2 common-guessing parameters). Parameter estimates for all 3 models are shown in Table 1. (Parameter estimates for the standard mixture model could not be obtained for some items, as discussed below, and these cells in Table 1 are empty.)

TABLE 1
*Parameter estimates of the single-state, mixture, and common-guessing models*

| | Single-State | | Mixture | | | | | Common-Guessing | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Item | $\hat{m}$ | $\hat{\sigma}$ | $\hat{m}_s$ | $\hat{\sigma}_s$ | $\hat{\rho}$ | $\hat{m}_G$ | $\hat{\sigma}_G$ | $\hat{m}_s$ | $\hat{\sigma}_s$ | $\hat{\rho}$ |
| 1 | 82.77 | 0.43 | | | | | | 82.77 | 0.43 | 0.00 |
| 2 | 79.27 | 0.51 | | | | | | 79.27 | 0.51 | 0.00 |
| 3 | 48.32 | 0.68 | | | | | | 54.13 | 0.60 | 0.12 |
| 4 | 32.27 | 0.67 | | | | | | 32.44 | 0.66 | 0.01 |
| 5 | 86.72 | 0.53 | | | | | | 89.53 | 0.50 | 0.04 |
| 6 | 42.54 | 0.55 | 44.41 | 0.52 | 0.08 | 18.77 | 0.88 | 43.68 | 0.53 | 0.04 |
| 7 | 72.69 | 0.40 | | | | | | 72.69 | 0.40 | 0.00 |
| 8 | 61.35 | 0.39 | | | | | | 61.37 | 0.39 | 0.00 |
| 9 | 83.80 | 0.37 | 84.25 | 0.36 | 0.01 | 4.74 | 0.52 | 84.32 | 0.36 | 0.01 |
| 10 | 155.26 | 0.34 | | | | | | 157.54 | 0.32 | 0.03 |
| 11 | 48.32 | 0.65 | | | | | | 48.74 | 0.65 | 0.01 |
| 12 | 81.77 | 0.54 | | | | | | 87.19 | 0.48 | 0.08 |
| 13 | 65.74 | 0.50 | 69.90 | 0.44 | 0.10 | 13.84 | 1.52 | 69.95 | 0.45 | 0.08 |
| 14 | 93.63 | 0.55 | 105.65 | 0.45 | 0.15 | 19.61 | 1.06 | 103.55 | 0.47 | 0.11 |
| 15 | 130.71 | 0.38 | 139.74 | 0.32 | 0.12 | 36.17 | 0.91 | 137.25 | 0.34 | 0.07 |
| 16 | 36.44 | 0.66 | 41.83 | 0.57 | 0.15 | 9.06 | 0.97 | 41.30 | 0.57 | 0.16 |
| 17 | 51.75 | 0.60 | 59.08 | 0.49 | 0.13 | 5.32 | 0.81 | 60.32 | 0.48 | 0.18 |
| 18 | 32.81 | 0.67 | 39.36 | 0.53 | 0.16 | 5.25 | 0.71 | 40.36 | 0.50 | 0.26 |
| 19 | 93.22 | 0.45 | 106.34 | 0.31 | 0.22 | 24.72 | 1.20 | 104.53 | 0.34 | 0.15 |
| 20 | 55.03 | 0.83 | 79.63 | 0.55 | 0.28 | 7.25 | 1.14 | 81.42 | 0.53 | 0.32 |
| 21 | 42.02 | 0.84 | 64.46 | 0.50 | 0.31 | 5.85 | 0.98 | 67.38 | 0.46 | 0.39 |
| 22 | 33.71 | 0.99 | 65.66 | 0.47 | 0.43 | 6.93 | 1.11 | 68.16 | 0.43 | 0.50 |
| 23 | 63.36 | 0.53 | 71.37 | 0.43 | 0.14 | 8.44 | 0.91 | 71.57 | 0.42 | 0.15 |
| 24 | 43.61 | 0.59 | 53.60 | 0.40 | 0.25 | 8.70 | 1.35 | 53.96 | 0.40 | 0.26 |
| 25 | 46.63 | 0.60 | 58.20 | 0.38 | 0.28 | 10.51 | 1.41 | 58.47 | 0.39 | 0.27 |

*Note.* The parameterization shown is the estimated median in seconds and the estimated standard deviation in ln(sec). For the mixture and common-guessing parameters, the subscripts S and G refer to the solution and guessing distributions, respectively. For the common-guessing model, $\hat{m}_G = 9.73$ and $\hat{\sigma}_G = 1.26$ for all items. Lines separate item sets (items 7, 8, 9, 23, 24, and 25 are not in a set).

## *Examining the Fit of the Models*

To evaluate the fit of the models, the root mean squared error (RMSE), was calcuated for each model. Additionally, the observed and predicted CDFs were plotted for each item and were inspected visually. Several representative items will be shown.

### *Root Mean Squared Error*

RMSE is the standardized difference between the observed and predicted function and can be thought of as a measure of dispersion centered on the predicted function. Larger values of RMSE indicate worse fits. RMSE is given by

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(F_{Ot} - \hat{F}_{Ot})^2} \quad (5)$$

where

- $t$ is the index of response times (RTs),
- $T$ is the number of observed response times,
- $F_{Ot}$ is the observed value of the CDF for the $t^{th}$ response time, and
- $\hat{F}_{Ot}$ is the predicted value of the CDF for the $t^{th}$ response time.

Figure 2 shows the RMSE for the single-state, standard mixture, and common-guessing mixture models for each item. The single-state model (Equation 2) fit the response-time distributions for most of the items on the first half of the test very well; on these items RMSE for the single-state model was small. On the last half of the test (beginning with item 12), the single-state model did not fit nearly as well; RMSE is markedly larger on these items. The standard mixture and common-guessing mixture models fit the response-time distributions for all items very well (when final estimates were available).[7] On the first 11 items, the single-state model performs as well or nearly as well as the standard mixture and common-guessing mixture models, but on the rest of the items, the difference in performance is much larger.

[7] The mixture model does not have final parameter estimates for some items early in the test because there were not enough rapid guesses for parameter estimates of the guessing distribution to stabilize on final values. For these items, Table 1 contains empty cells for the mixture model, and Figure 2 has no value of RMSE. (The guessing-distribution parameters of the common-guessing model are estimated from only those items that are not adequately described by the single-state model; as long as $\hat{\rho} \neq 0$ for at least one item, the parameter estimates for the common-guessing model will stabilize on final values. Thus there are no empty cells in Table 1 for the common-guessing model, and there are RMSE values for the common-guessing model for all items.)

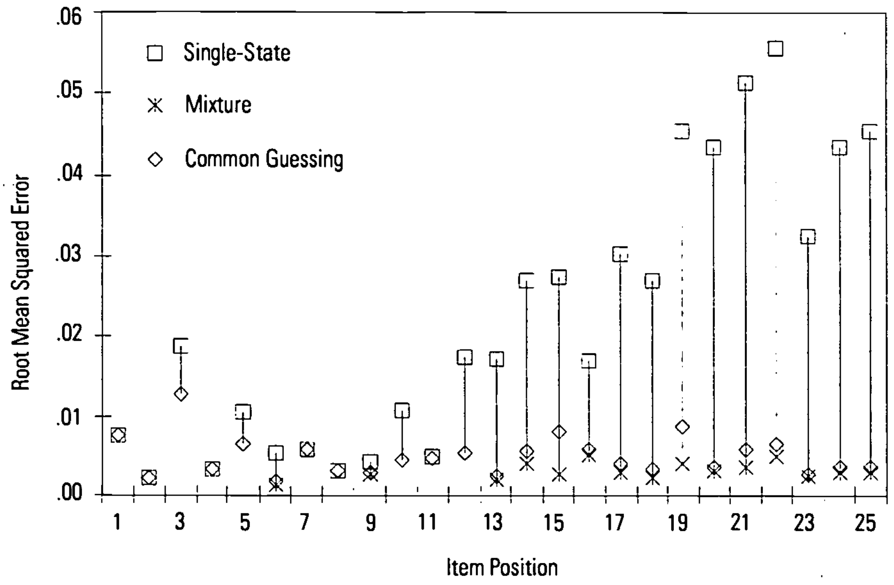

FIGURE 2. *Root mean squared error for the single-state, mixture, and common-guessing models.*

*Visual Inspection of the Cumulative Distribution Functions (CDFs)*

For a more detailed look at how the various models fit the data, the CDFs for several representative items are shown in Figures 3-6. Item 4, shown in Figure 3, is typical of items on the first half of the test. On this item, the observed CDF is barely distinguishable from the CDFs predicted from the single-state and common-guessing mixture models. The common-guessing mixture model makes predictions that are virtually identical to those of the single-state model for item 4 because $\hat{\rho}$ is so small. Parsimony dictates using the single-state model on these items; guessing is not a factor for early items.

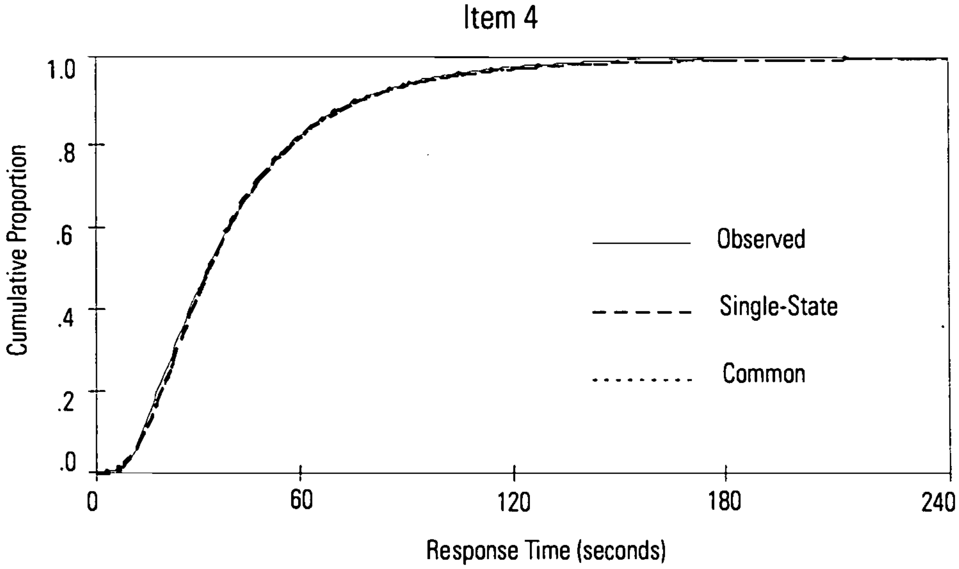

FIGURE 3. *Cumulative distribution function of observed response times for item 4 and predicted values of the single-state and common-guessing models.*

Items on the last half of the test were not fit as well by the single-state model, as indicated by the RMSE values in Figure 2. The response-time distributions for items on the last half of the test were weakly bimodal. The second mode was formed by an unusual number of fast responses. On all of these items, the single-state model underpredicted the number of fast responses, as shown for item 17 in Figure 4. As the proportion of rapid guesses increased (according to the standard mixture model), the fit of the single-state model became worse (as in Figure 5, item 22). On these two items, as well as all other items on the last half of the test, the standard mixture model fit the entire response-time distribution very well (also indicated by the RMSE values).

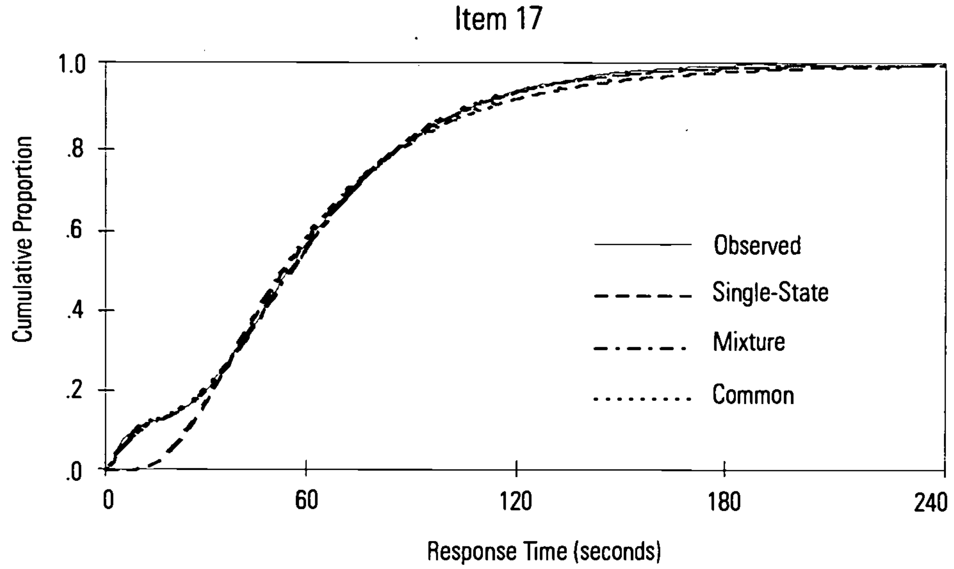

FIGURE 4. *Cumulative distribution function of observed response times for item 17 and predicted values of the single-state, mixture, and common-guessing models.*

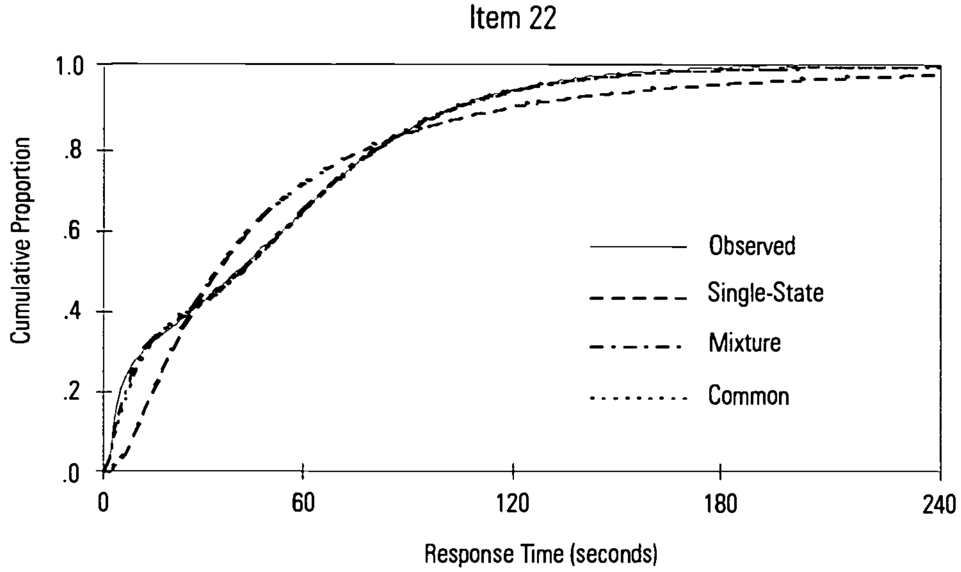

FIGURE 5. *Cumulative distribution function of observed response times for item 22 and predicted values of the single-state, mixture, and common-guessing models.*

### *Addressing the Hypotheses*

*Is Rapid-Guessing Behavior the Same Across Items?*

Both the standard mixture model and the common-guessing mixture model fit very well for most of the items on which the single-state model failed (as in Figures 4 and 5). The common-guessing mixture model fit items that were the first in a set (items 10, 15, and 19)[8] slightly less well than other items. As shown in Figure 2, the RMSE for the common-guessing mixture model on items 15 and 19 are slightly larger than that of the standard mixture model (although still quite small). On these items, the common-guessing mixture model slightly overpredicted the number of fast responses (see, for example, item 19 in Figure 6; the difference was just large enough for the lines in the graph to be distinguishable), but the fit was still quite good. Thus it appears that the rapid-guessing response-time distribution is essentially the same across items.

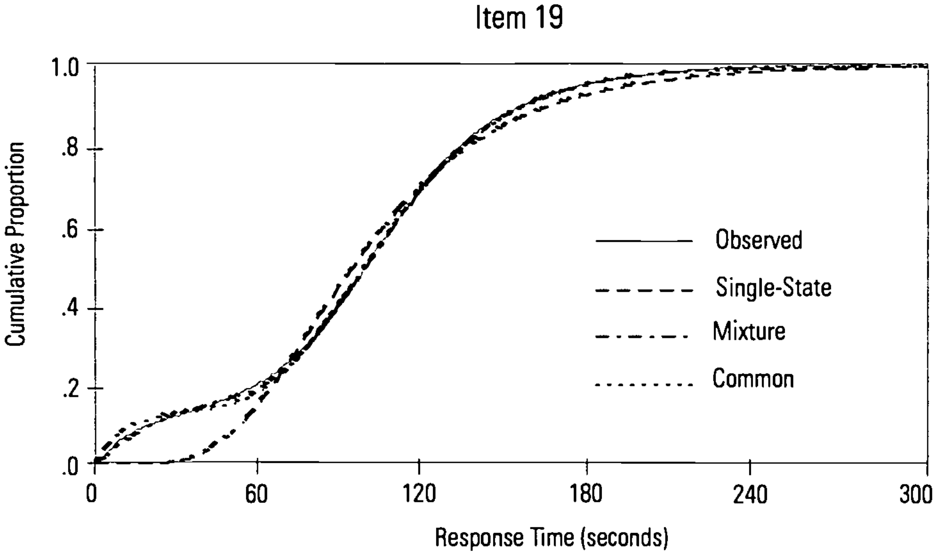

FIGURE 6. *Cumulative distribution funtion of observed response items for item 19 and predicted values of the single-state, mixture, and common-guessing models.*

*Does Rapid-Guessing Behavior Increase Toward the End of the Test?*

To address whether rapid-guessing behavior increases during the test, the patterns of $\hat{\rho}$'s from the standard mixture and common-guessing mixture models (shown in Table 1) were inspected. The $\hat{\rho}$'s obtained from the fits of the standard mixture and common-guessing mixture models tend to increase across item position. However, there is also some evidence that item set influences the trend: $\hat{\rho}$ for the first item of a set tends to be somewhat lower than the $\hat{\rho}$ for the previous item, and the $\hat{\rho}$'s increase within the set. This trend is also evident in the RMSE results in Figure 2. (The $\hat{\rho}$ for item 3 from the common-guessing mixture model is uncharacteristically high for that early in the test. Notice also that the RMSE values for both the single-state and common-guessing mixture models on item 3 are larger than the values for nearby items. There was clearly something peculiar about the response times on this item, and this was reflected in $\hat{\rho}$.)

[8] Item 1 was also the first of a set, but there was no rapid-guessing behavior on this item, as shown in Table 1.

*Is Accuracy an Increasing Function of Response Time?*

As shown in Figure 7, accuracy starts near chance[9] for fast response times, then increases as more time is spent processing an item, up to a plateau that varies by item. (The upper tails in Figure 7 have been truncated for each item because there are too few test takers to obtain stable estimates; between 2% and 6.25% of the test takers were trimmed from each item.)

To relate the accuracy function to the underlying distributions specified by the two-state model, the point at which the hypothesized underlying rapid-guessing and solution distributions intersect was calculated numerically for both the standard mixture and common-guessing mixture models. These points are represented as vertical lines in Figure 7. The lines intersect the increasing portion of the accuracy functions, as would be predicted by a two-state mixture model because the accuracy at these points should be an even mixture of the accuracy rate obtained by rapid-guessing behavior and the accuracy rate obtained through solution behavior. Notice that the intersecting lines for the two forms of the model (standard and common-guessing) are similar, as would be expected because the parameters of the models are similar (e.g., see Table 1).

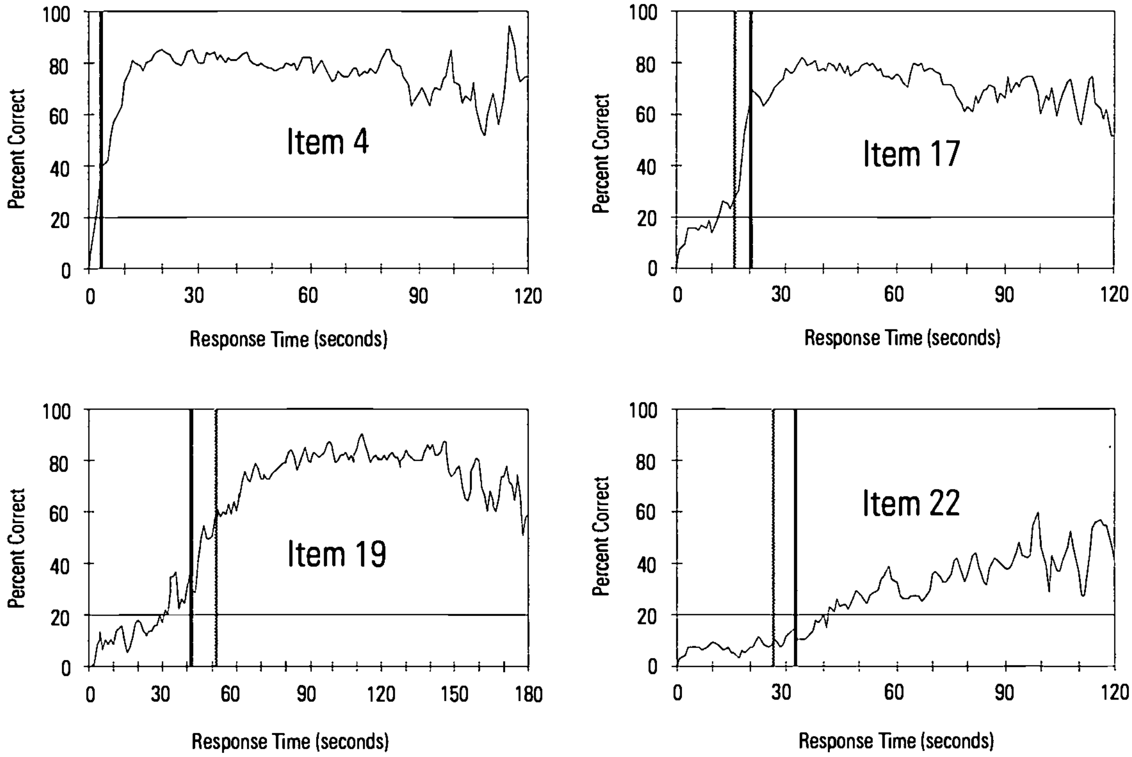

FIGURE 7. *Accuracy as a function of response time for the representative items. The upper tail for each item is truncated (2% to 6.25% of the responses were trimmed). Vertical lines show where the hypothetical underlying guessing and solution distributions cross for the mixture model (dotted line) and the common-guessing model (solid line).*

BEST COPY AVAILABLE

[9] Actually, accuracy is often below chance for rapid guesses. More will be said about this finding in the Discussion.

*The Amount of Speededness on the Test*

To determine the amount of speededness on the test, the proportion of rapid-guessing behavior ($\hat{\rho}_i$) was combined with the proportion of test takers who did not reach each item. To do this, $\hat{\rho}_i$ (from the common-guessing mixture model) was adjusted so that it is the proportion of rapid-guessing behavior in the entire test taker population, rather than the proportion of rapid-guessing behavior among those who reached the item. These adjusted values, along with the proportion of test takers who did not reach the item, are shown in Figure 8 as a stacked bar chart. As shown, about 15% of the test takers did not reach the last item. In the past, this number may have been used to describe the amount of speededness on the test. When rapid-guessing behavior is also considered, a much higher proportion of test takers affected by speededness is revealed. It is clear from Figure 8 that speededness affects all items on the last half of the test, and the proportion of test takers who are affected is quite substantial (up to .53). The relationship between rapid-guessing behavior and item sets that was revealed in Table 1 is also evident in Figure 8; rapid-guessing behavior (and hence speededness) increases within a set, then decreases when a new set (or a nonset) starts. The item most affected by speededness was item 22 which was the last item of the last set, not item 25 which was the last item on the test.

Test-level indices of speededness could also be derived. For example, one could calculate the percentage of items on which more than 10% (or some other percentage) of the test takers were speeded (engaged in rapid-guessing behavior or did not reach the item). Alternately, the maximum percentage of speeded behavior on any one item could be calculated. Additionally, the mean percentage of speeded behavior across items could be calculated to provide an indication of test-level speededness.

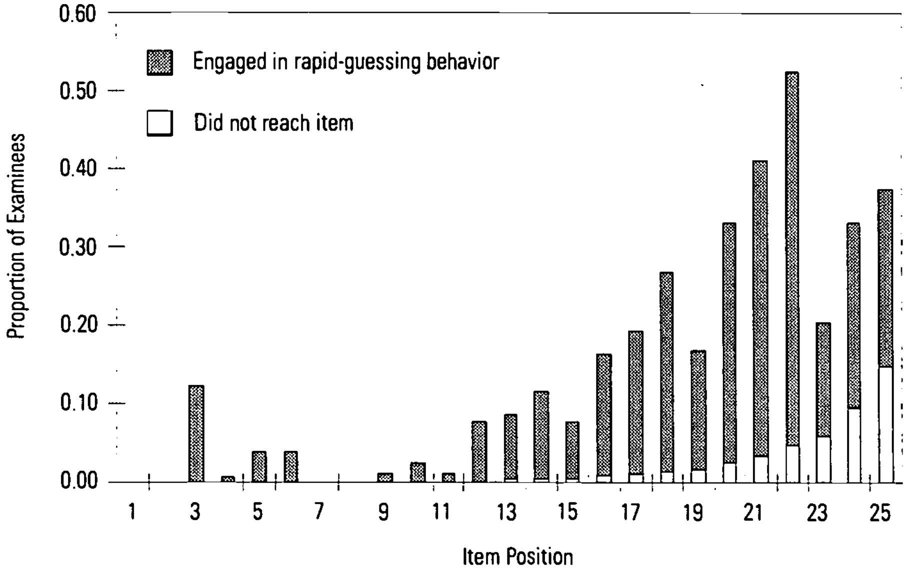

FIGURE 8. *Proportion of examinees who were affected by speededness on each item, separated by those who did not reach the item and those who engaged in rapid-guessing behavior. (The proportion of examinees engaging in rapid-guessing behavior was adjusted to represent the proportion in the total examinee population, rather than the proportion of those who reached the item.)*

## Discussion

Strict time limits may cause some test takers to rapidly answer items without fully considering the items, especially if no points are subtracted from the test score for wrong responses, as on the LSAT. This is a wise strategy for test takers to use, but it causes difficulties when speededness is being measured to determine if speed was having too large an influence on test taker performance. Traditional speededness indices depend on "unreached" items, but when test takers engage in rapid-guessing behavior, the items are not unreached. More recent approaches to speededness depend on the drop in accuracy at the end of the test (e.g., Yamamoto, 1995). The latter approach is better because it incorporates guessing, but a more direct investigation of speededness would focus on the amount of time spent on each item.

Response times for individual items are not available on large-scale operational paper-and-pencil tests, but for computer-administered tests, response times can be collected easily and unobtrusively. The present study used response-time data from a computer-administered version of the GRE to develop a way to detect rapid-guessing behavior (and hence speededness) on an operational, high-stakes examination.

It was clear from inspecting the observed probability density functions that items on the last half of the test had more fast responses than would be expected if time had not been a factor. Many test takers were responding in less than 10 seconds on these items, whereas the median response time was a minute or more. Test takers who responded within several seconds clearly did not fully consider the item. At most, they may have scanned the item and answer choices for key words.

In a more formal, model-based fit of response times, a single-state model was unable to account for the number of fast response times on items on the last half of the test, although it generally fit well for items on the first half of the test. Consistent with this finding, parameter estimates for a two-state standard mixture model were not obtained for most items on the first half of the test because $\rho$, the proportion of responses in the rapid-guessing distribution, was too small. Parameter estimates for the standard mixture model were obtained for all items on the last half of the test where rapid guesses were more common.

Results suggest that rapid-guessing behavior is independent of item content and is a function of item position. The underlying rapid-guessing distribution appears to be the same for all items, regardless of item content, difficulty, length, or other item characteristics, and the pattern of $\hat{\rho}$'s suggests that rapid-guessing behavior increases during the test, although set membership alters the pattern. Rapid-guessing behavior appears to increase within a set (this is most noticeable when looking at the $\hat{\rho}$'s from the common-guessing mixture model). When a new set (or a nonset) starts, rapid-guessing behavior drops. This suggests that test takers may "give up" during a set but try again when a new set (or a nonset) starts. A test with no set-bound items should have strictly increasing $\hat{\rho}$'s.

Finally, accuracy as a function of response time provides additional support for a two-state mixture model. These functions show low accuracy levels for fast response times, rise to higher levels of accuracy for slower response times, and reach a plateau for which increasing time does not increase accuracy; the location of the increasing portion of the function corresponds to the rapid-guessing and solution distributions predicted from the model.

The accuracy functions provided some unexpected information as well—accuracy for the "fast" distribution was not at chance (0.2), as would be expected if test takers were simply guessing randomly. On most items, accuracy for fast response times is below chance. On a few items (10 and 15—both the first of a set), accuracy for the fast (rapid-guessing) distribution is actually above what would be expected by chance, but below the accuracy associated with the "slow" (solution-behavior) distribution. Taken together with the response-time results, these accuracy functions suggest that test takers are engaging in two types of behavior, but the behavior associated with the "fast" distribution is not necessarily *random* guessing. Test takers, as a group, may be engaging in a form of systematic guessing (e.g., "always respond A" or "pick the longest response option") or may quickly seek key words (that often lead to distractors) without carefully analyzing the item and answer options. Such a strategy would result in below-chance performance on most items and above-chance on a selected few. These possibilities could be tested by examining the distribution of responses (A, B, C, D, or E) as a function of response time, but we do not currently have access to the response-selection data for the current test, so these possibilities could not be explored in the present study. In any case, these findings provide an important direction for future research.

The point of modeling response times with a two-state mixture model (and verifying that the second underlying distribution was consistent with what we expect rapid-guessing behavior to be) was to detect an additional type of speeded behavior (i.e., rapid-guessing behavior) which could then be combined with the traditional indication of speededness (i.e., not reaching the item) to provide a more accurate measure of the total amount of speededness in the test. The proportion of test takers who did not reach each item monotonically increased up to about .15 by the last item, so even by traditional standards this test seems somewhat speeded. Using the mixture-model approach, we found even stronger evidence of speededness. Fully half of the items showed evidence that test takers switched from a solution-based strategy to a rapid-guessing strategy, and the percentage of test takers who switched to a rapid-guessing strategy (indicating speededness) ranged up to 50%.

Displaying speededness item-by-item allows a more in-depth investigation of the phenomenon than a summary statistic would allow. The item-level approach was able to show that speededness increased within a set. In related work, Schnipke (1995) found that speededness may affect particular item types more than others. She found that only reading comprehension items on the verbal section of the GRE showed evidence of rapid-guessing behavior, and these items were located in the middle of the section. Inspecting all items for evidence of speededness provides information that would be lost to summary statistics.

### *Implications*

Time limits are useful for administrative reasons and scheduling purposes, and if the time limits are liberal enough, there may be few negative consequences. Determining if the time limit is adequate is crucial. If nearly all test takers are able to consider fully most of the items (i.e., engage in solution behavior), the time limit has little effect on test scores, in which case the degree of speededness is very small and can be ignored. This was not the case on the test analyzed in the present study; it was largely affected by the time limits.[10]

As LSAC considers computerizing the LSAT, the benefits of obtaining item response times should be considered. With response times available, methodology such as that developed in the present study can be used to detect rapid-guessing behavior. This is not possible on paper-and-pencil administered tests. How large an effect time limits have on the LSAT is not currently known because we cannot identify rapid-guessing behavior.

If a time limit is used on an LSAT computerized test, it is possible to determine if the time limit has detrimental effects on test taker behavior. Both the number of test takers who do not finish the test and the response-time distributions should be considered. If most test takers reach all items and there are no unusually large groups of fast responses in the probability density functions, and if the single-state model fits the distribution function for all items (including any fast response times), then it would seem to be the case that the time limit was not having a detrimental effect on test taker behavior.

The proportion of test takers who engaged in rapid-guessing behavior on each item, $\hat{\rho}$, could be used as an item-level index of rapid-guessing behavior (one kind of speededness) and should be added to the proportion of test takers who did not reach the item for a complete item-level measure of speededness. This would be a more appropriate approach to measuring the speed component in speeded tests than techniques currently in use.

[10] The version of the GRE used in the present study was not typical of other administrations because the computerized version of the GRE is now administered adaptively and with a different time limit. Thus our conclusions about the amount of speededness in the GRE section analyzed in the present study should not be generalized to other administrations of the GRE. Our goal is not to show how speeded the GRE is, but rather it is to develop a new method for detecting rapid-guessing behavior in general.

The present methodology could be used in its current form on nonadaptive computer-administered tests (such as the test used in the present study) and possibly on tests which are composed of sequentially-administered testlets (e.g., Lewis & Sheehan, 1990; Luecht, Nungester, & Hadadi, 1996; Reese & Schnipke, 1996; Sheehan & Lewis, 1992). (In the sequentially-administered testlet design, the technique could be applied most directly if the testlets were separately timed; each testlet could be analyzed separately).

The methodology cannot be directly applied to computerized adaptive testing (CAT) data, however. In CAT, items appear in different locations in the test, absolute time limits do not necessarily apply, and test taker ability is generally more homogeneous for a given item (especially items at extreme difficulty levels). These factors add a level of complication that the current methodology does not cover. (Homogeneity of test taker ability is a factor if response times are related to test taker ability.) Additional research is needed to adapt the methodology to CAT data and also to predict speededness on CATs.

## Conclusions

Response times have only recently become available in operational tests, and the current work is a first step toward understanding an important aspect of response-time data. The current work also demonstrates the importance of exploring the shapes of response-time distributions instead of focusing on summary statistics. There is a great deal of response-time research that follows the current work. For instance, test takers would presumably be less likely to rapidly guess on items if there was a penalty for incorrect responses, and future research could investigate this topic. Future research could also investigate the possibility of rescoring tests after removing rapid guesses (see Potenza & Stocking, 1994, for similar work on rescoring tests with "flawed" items).

## References

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

Gulliksen, H. (1950). *Theory of mental tests.* Hillsdale, NJ: Lawrence Erlbaum.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer-Nijhoff.

Lewis, C., & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement, 14,* 367-386.

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.

Mollenkopf, W. G. (1950). An experimental study of the effects on item-analysis data of changing item placement and test time limit. *Psychometrika, 15,* 291-317.

Norušis, M. J. (1994). *SPSS Advanced Statistics*$^{TM}$ 6.1. Chicago, IL: SPSS.

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

Potenza, M. T., & Stocking, M. L. (1994). *Flawed items in computerized adaptive testing* (Report No. 94-6). Princeton, NJ: Educational Testing Service.

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). *Assessing speededness in computer-based tests using item response times.* Unpublished doctoral dissertation, Johns Hopkins University, Baltimore.

Schnipke, D. L. (1996, April). *How contaminated by guessing are item-parameter estimates and what can be done about it?* Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Sheehan, K., & Lewis, C. (1992). Computerized mastery testing with nonequivalent testlets. *Applied Psychological Measurement, 16,* 65-76.

Swineford, F. (1956). *Technical manual for users of test analysis* (Statistical Report 56-42). Princeton, NJ: Educational Testing Service.

Townsend, J. T., & Ashby, F. G. (1983). *The stochastic modeling of elementary psychological processes.* Cambridge, England: Cambridge University.

Yamamoto, K. (1995). *Estimating the effects of test length and test time on parameter estimation using the HYBRID model* (TOEFL Technical Report TR-10). Princeton, NJ: Educational Testing Service.

# NOTICE

## Reproduction Basis