DOCUMENT RESUME

ED 383 742 TM 023 219

AUTHOR Schnipke, Deborah L.
TITLE Assessing Speededness in Computer-Based Tests Using Item Response Times.
SPONS AGENCY Educational Testing Service, Princeton, N.J.; Graduate Record Examinations Board, Princeton, N.J.
PUB DATE Apr 95
NOTE 32p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 19-21, 1995).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Computer Assisted Testing; *Educational Assessment; Evaluation Methods; *Guessing (Tests); *Item Response Theory; Responses; *Test Items; Test Reliability; *Timed Tests; Time Factors (Learning)
IDENTIFIERS Accuracy; Graduate Record Examinations; *Speededness (Tests)

ABSTRACT

Time limits on tests often prevent some examinees from finishing all of the items on the test; the extent of this effect has been called the "speededness" of the test. Traditional speededness indices focus on the number of unreached items. Other examinees in the same situation rapidly fill in answers in the hope of getting some of the items right by chance. Those examinees who do not have unanswered items are not included in traditional measures of speededness. To obtain an accurate measure of speededness, however, those who guess rapidly also need to be included in the estimate of speededness. They will have very fast response times, and the responses will be at or near chance levels of accuracy. Therefore, item response times, in conjunction with accuracy rates, can be used to identify these examinees and provide a more rigorous measure of speededness than has previously been available. Analyses based on item response time distributions in the analytical section of the Graduate Record Examinations (GRE) General Test for over 17,000 examinees indicate that many examinees do guess rapidly on items and that the test is more speeded than traditional measures of speededness indicate. Six tables and seven figures present analyses results. (Contains eight references.) (Author)

Assessing Speededness In Computer-Based Tests Using Item Response Times

Deborah L. Schnipke
Johns Hopkins University

A paper presented at the annual meeting of the
National Council on Measurement in Education
April, 1995, San Francisco, CA

Assessing Speededness of Computer-Based Tests Using Item Response Times[1]

Abstract: Time limits on tests often prevent some examinees from finishing all of the items on the test; the extent of this effect has been called the "speededness" of the test. Traditional speededness indices focus on the number of unreached items. Other examinees in the same situation rapidly fill in answers in the hopes of getting some of the items right by chance. These examinees will not have unanswered items and therefore are not included in traditional measures of speededness. To obtain an accurate measure of speededness, however, examinees who rapidly guess on items also need to be included in the estimate of speededness. Examinees who rapidly guess on items will have very fast response times, and the responses will be at or near chance levels of accuracy. Therefore, item response times, in conjunction with accuracy rates, can be used to identify these examinees and provide a more rigorous measure of speededness than has previously been available. Analyses based on item response time distributions in the analytical section of the Graduate Record Examinations (GRE) General Test indicate that many examinees do rapidly guess on items, and the test is more speeded than traditional measures of speededness indicate.

In testing, a distinction is made between tests that measure power and tests that measure speed (Gulliksen, 1950). In a pure power test, the items range in difficulty and there is no time limit. The goal is to measure how accurately the examinees can answer the items. In a pure speed test, the items are very easy and the time limit is very strict. The goal is to measure how quickly the examinees can answer items. In reality, most tests contain both speed and power components, and these tests are called speeded tests. Speeded tests usually result from administering a power test with a time limit, a practice that is usually required when the test is group-administered.

---

[1] Support for this research was provided by Educational Testing Service (ETS) and the Graduate Record Examinations (GRE) Board through the GRE Graduate Research Assistantship in Psychometrics Program. The points of view and opinions expressed in this paper do not necessarily represent official ETS or GRE Board position or policy.

Speededness is a problem for both classical test theory and item response theory (IRT). Split-half reliabilities are spuriously high for speeded tests (Gulliksen, 1950), and item position influences item indices (Anastasi, 1988). Unidimensional IRT implicitly assumes that the test is unspeeded; speed would be another dimension (Hambleton & Swaminathan, 1985). Items which were speeded for an examinee should be ignored when calculating the examinee's ability if the speeded responses can be identified (Lord, 1980). When estimating IRT item parameters on a simulated speeded test, the **a** and **b** (discrimination and difficulty) parameters were overestimated and the **c** (guessing) parameters were underestimated for the items toward the end of the test (Oshima, 1994).

Because speededness is a problem for test theory, one might argue against having time limits on tests. However, having a time limit does not necessarily mean that the test is speeded. If nearly all the examinees are able to fully consider and answer almost all of the items, the time limit is not important. Thus, if the degree of speededness in a test is very small, it can be ignored, and the test may be considered to be a pure power test, even when administered with a time limit.

Speededness has traditionally been defined as the extent to which some examinees are unable to finish all items on the test and has been measured by the percentage of examinees who do not reach a certain number of items (e.g., all items, or 75% of the items). The problem with this definition is that if examinees rapidly guess on some items as time expires in an attempt to get the items right by chance (rather than leave the items blank), they will have fewer or no unreached items; the test will not appear to be speeded for these examinees. However, this behavior is also the result of speededness, and to accurately measure speededness, these rapid guessers must be included in the estimate. A

better definition of speededness is the extent to which some examinees are *disadvantaged* by the time limit on a test, relative to other examinees. Examinees who rapidly guess on items, as well as examinees who do not reach items, are disadvantaged by the time limit and are included in the estimate of speededness that is developed in the present study.

Figure 1 shows the behavior of a typical speeded examinee. The examinee's standardized natural logarithm of response time is shown for each item. The natural logarithm is used because the response time distributions are positively skewed, and the natural logarithm transformation creates a more normal distribution, as will be shown later. The transformed response times were then standardized, item by item, to control for item differences (e.g., long items take longer, on average, to answer than short items). If an examinee responded at the mean speed on an item, the standardized natural logarithm of the response time (standardized ln(RT), for simplicity) would be 0. Positive values of standardized ln(RT) indicate that the examinee responded slower than the average examinee, and negative values indicate that the examinee responded faster than average.

The examinee depicted in Figure 1 responded more slowly than most examinees on the first few items (items 1-7), then sped up and responded at about the same speed as most examinees on the next several items (items 8-13). The examinee did not finish the rest of the test. This examinee was slow, but accurate: of items 1-12, all were answered correctly except item 9. The examinee did not respond to item 13 (i.e., it was omitted), and the rest (items 14-25) were unreached (i.e., the items were never displayed on the screen). Traditional speededness indices would identify this person as speeded.

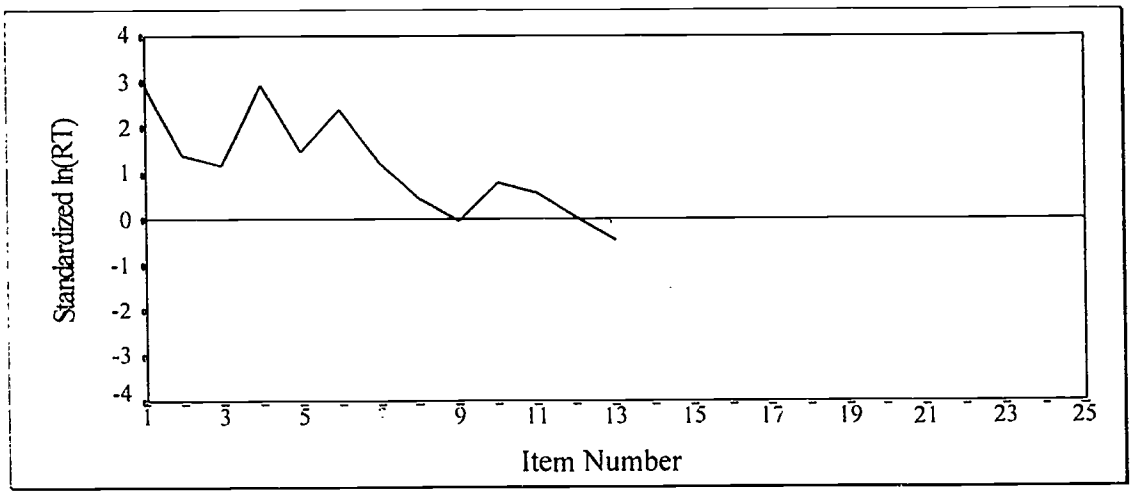

Figure 1: Standardized ln(RT) across items for a typical speeded examinee. (The examinee did not finish all items on the test.)

Not finishing all the items on a test is not the only type of speeded behavior. Rapid guessing on items can also be due to speededness. In *rapid-guessing behavior*, the examinee responds rapidly to items as time expires, and accuracy will be at or near chance because the examinee is not fully considering the items. The examinee may skim the item briefly for keywords, but the examinee does not completely read the item when engaging in rapid-guessing behavior. Consequently, item characteristics may have little effect on response times.

In contrast, in *solution behavior*, the examinee actively tries to determine the correct answer to every item. The examinee reads each item carefully and fully considers the answer. Accuracy will depend on item difficulty and other item characteristics and on the examinee's ability. Thus, response times resulting from solution behavior will be

better predicted by item characteristics than response times that result from rapid-guessing behavior.

In a completely non-speeded test, all examinees would presumably engage in solution behavior on all items. Thus, rapid-guessing behavior by an examinee implies that the test is speeded for that examinee. An accurate measure of test speededness requires that examinees engaging in rapid-guessing behavior be identified. This can be done by examining the response time distributions for fast inaccurate responses.

Figure 2 shows the behavior of an examinee who switched from solution behavior to rapid-guessing behavior. At the beginning of the test, the examinee responded at a slower rate than most examinees (from .5 to 1.5 standard deviations above the mean), but after item 13 (when the examinee had only two minutes remaining), the examinee suddenly started responding much faster (from -.7 to -3.9 standard deviations below the mean). The examinee's accuracy also changed – from 77% accuracy on the first 13 items to 25% accuracy on the last 12 items. The examinee engaged in solution behavior on the first half of the items, but on the last half of the items (and with very little time left), the examinee engaged in rapid-guessing behavior. Traditional measures would not include this examinee in the estimate of speededness, but the test is clearly speeded for this examinee.

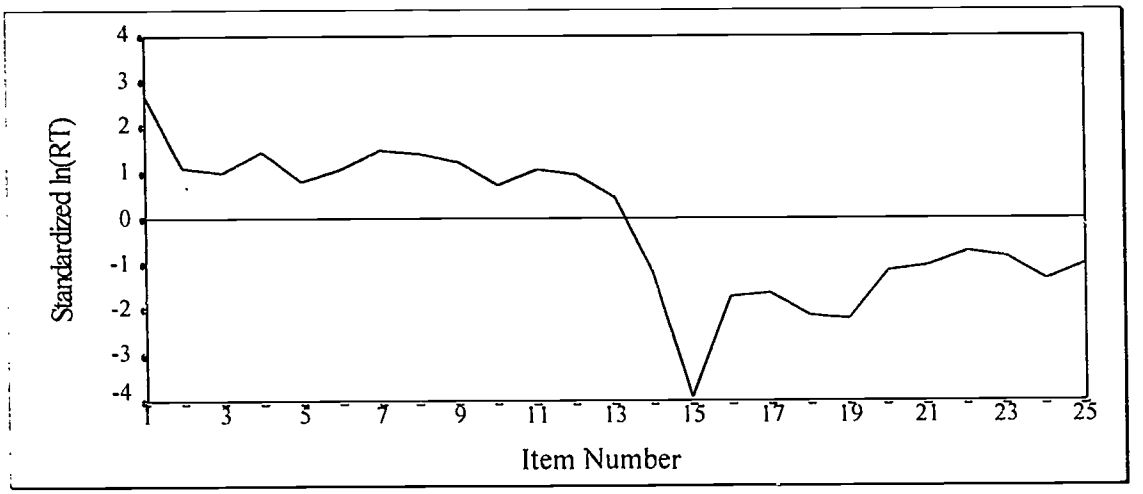

Figure 2: Standardized ln(RT) across items for a non-typical speeded examinee (non-typical because the examinee finishes the test by rapidly guessing on items at the end of the test).

On a speeded test, there are likely to be both non-speeded items and speeded items. It is hypothesized that on the non-speeded items, response times will be predicted well by item characteristics because all of the examinees engage in solution behavior. On the speeded items, there will be many fast, inaccurate responses resulting from rapid-guessing behavior, as well as responses arising from solution behavior. It is hypothesized that the response times on speeded items will not be predicted as well by item characteristics because of the rapid-guessing behavior. Removing responses due to rapid-guessing behavior should allow item characteristics to be better predictors of response times on the speeded items because only solution behavior will remain. It is also hypothesized that accuracy (right or wrong) will be a better predictor of response time on speeded items than on non-speeded items because wrong responses are more likely to be

fast responses on speeded items. Finally, it is hypothesized that the non-speeded items will be more likely to occur at the beginning of the test and the speeded items at the end of the test, although item type is likely to modulate this pattern.

METHOD

In order to find rapid-guessing behavior, a test that is known to be somewhat speeded was needed. In order to obtain item response times, a computer-administered test was needed. One of the Analytical sections from a computer-based Graduate Record Examinations General Test (the GRE-CBT) was used because it met both needs.

Examinees and Test Forms

The GRE-CBT was administered to 17,415 students in the 1992-1993 academic year. All 17,415 examinees chose to take the computer version of the GRE rather than the regular paper-and-pencil version. The GRE-CBT was administered by Sylvan Kee Centers which are operated by Sylvan Learning Systems, Inc. Three test forms (J, N, and O) were administered to examinees. These forms were originally paper-and-pencil forms that were converted for computer delivery. Form J was administered to 7,218 examinees, Form N to 7,001 examinees, and Form O to 3,196 examinees. Because Form J had the largest sample size, it was used for data analyses.

Test Sections

As in the paper-and-pencil GRE, the GRE-CBT had seven sections (subtests): two verbal sections of 38 items each, two quantitative sections of 30 items each, two

analytical sections of 25 items each, and one non-operational (non-scored) section. Items were, in general, arranged in order of ascending difficulty.

Most items on the analytical sections were arranged in sets that referred to a common stimulus. In Analytical Section 1, which will serve as the data set for the present study, there were 4 sets. Items 1-6 were in the first set, items 7, 8, and 9 were not in a set, items 10-14 were in the second set, items 15-18 were in the third set, items 19-22 were in the fourth set, and items 23, 24, and 25 were not in a set. For items in sets, the common stimulus was presented on the left half of the computer screen, and the items were presented one-at-a-time on the right side of the screen.

Test Administration

Items were presented singly on a computer screen. Responses were made by clicking on the desired response with the mouse (a hand-held pointing device that controls the position of the cursor on the screen). Examinees had 32 minutes to complete each section[2], and an optional 10-minute break was offered between Sections 3 and 4. Unlike some computer-administered tests, the GRE-CBT allowed examinees to omit items and return to previously viewed items. Examinees could change their previous answe. , and they could skip over items without ever seeing them.

Tutorials were presented on the computer to all examinees before the actual test started. The tutorials taught the examinees how to use the computer interface. There was

---

[2] The GRE-CBT allowed 32 minutes per section instead of 30 minutes as in the paper-and-pencil version to allow for the time it took to refresh the computer screen between items (generally less than 2 seconds per item).

no time limit on the tutorials, and examinees could repeat tutorial screens for additional practice.

RESULTS

Traditional speededness indices indicate that the analytical measure of the GRE is somewhat speeded. Several traditional indices are shown in Table 1 for the analytical sections in the present data set (the 1992-1993 GRE-CBT). The percentage of examinees answering the last item is the most commonly used measure of speededness. As shown in Table 1 under the heading "%Examinees Reaching Last Item," not all of the examinees reached the last item in either analytical section. In a strictly unspeeded test, the percentages would be 100%.

Table 1: Speededness Indices Based on the Number of Unreached Items

Test Section	%Examinees Reaching Last Item	ETS Rule-of-Thumb	
		%Examinees Reaching 75% of the Items	%Items Reached by 80% of the Examinees
Analytical 1	85.3	98.4	100.0
Analytical 2	78.3	96.8	96.0

Requiring all examinees to reach all items in order to consider a test completely unspeeded is a very strict rule. Perhaps the rule can be relaxed such that if all of the examinees reach most of the items, or most of the examinees reach all of the items, the test may be considered unspeeded. Defining "most" is then necessary. Swineford (1956) did just that in developing the ETS "rule of thumb." She stated that if all examinees

reach at least 75% of the items and all of the items are reached by at least 80% of the examinees, the test may be considered unspeeded. As shown in Table 1 under "%Items Reached by 80% of the Examinees," all of the items (100%) were reached by at least 80% of the examinees in Analytical 1, but in Analytical 2 only 24 of the 25 items (96%) were reached by at least 80% of the examinees. Also, all of the examinees did not reach at least 75% of the items in either section, as shown under "%Examinees Reaching 75% of the Items." Therefore, using the ETS rule-of-thumb, neither of the sections may be considered unspeeded.

Both analytical sections show evidence of speededness (according to traditional measures), thus either could be used in the present study. Analytical 1 was selected.

Data Cleaning[3]

The GRE is a high-stakes exams on which examinees are very motivated to do well. However, it was still necessary to remove examinees who did not take the test seriously. Thus, before any analyses were performed, two data cleaning techniques were used.

The first technique was to find extreme outliers in the data set. All examinees who spent 1000 or more seconds (16.67 minutes) on any one item were identified for investigation. Six examinees were identified.

---

[3] Analytical 1 was selected, then the data were cleaned. The indices presented in Table 1 are based on the cleaned data.

Three of the six examinees who were identified as extreme outliers spent all of the time on Item 1 but did not respond to it. Items 2 though 25 were not seen by these 3 examinees.

The fourth examinee spent 34 seconds on Item 1, answered it incorrectly, spent an average of 4 seconds (SD = 3 seconds) on Items 2 through 11 without answering any of them, then spent the remainder of the time (1835 seconds, or 30.58 minutes) on Item 12 but did not answer it. Examinee 4 did not see Items 13 through 25.

The fifth examinee spent 1482 seconds (24.7 minutes) on Item 1, did not answer it, then spent an average of 0.71 seconds (SD = 0.8 seconds) on the remaining items (2 through 25) without answering any of them.

The sixth examinee spent an average of 4 seconds (SD = 4.6 seconds) on the first 24 items, answered all of them and got 5 of them right. The examinee then spent the remainder of the time (1799 seconds, or 29.98 minutes) on the last item and answered it incorrectly.

All six of the examinees described above were removed from the data set because they were clearly not trying to do their best.

The second data cleaning technique focused on short overall section times. Examinees had the option of exiting a particular section at any time. Therefore, not all examinees spent all 1920 seconds (32 minutes) on the current section, Analytical 1. Examinees who spent very little time on the section and responded at or below chance were also removed from the data set. All examinees who spent less than 380 seconds (6.33 minutes) on the entire section were responding at chance. Therefore, all examinees

who spent less than 380 seconds on the section were removed from the data set (78 examinees).

Item Response Time Distributions

The purpose of the present study is to identify examinees engaging in rapid-guessing behavior, and this is done using item response times. Figure 3 shows response time distributions for several items in Analytical 1. Response time is plotted on the horizontal axis, and the number of examinees who responded at each level of response time is plotted on the vertical axis. The graphs are stacked bar charts (at each response time level, right responses are stacked on top of wrong responses, not behind them). Items that appear later in the test have more examinees responding very quickly (rapid-guessing), as expected in a speeded test.

Item 1 (Figure 3) shows no evidence of rapid-guessing behavior. The distribution is positively skewed (some examinees responded very slowly). Item 1 was very easy; there are almost no wrong responses.

Items 5 and 9 (Figure 3) show evidence of a small amount of rapid-guessing behavior, as seen by the fast wrong responses which stand out to the left of the rest of the response time distribution. Overall, the distributions are positively skewed with only a small amount of rapid-guessing behavior, and this is the case in general for the items on the first half of the test.

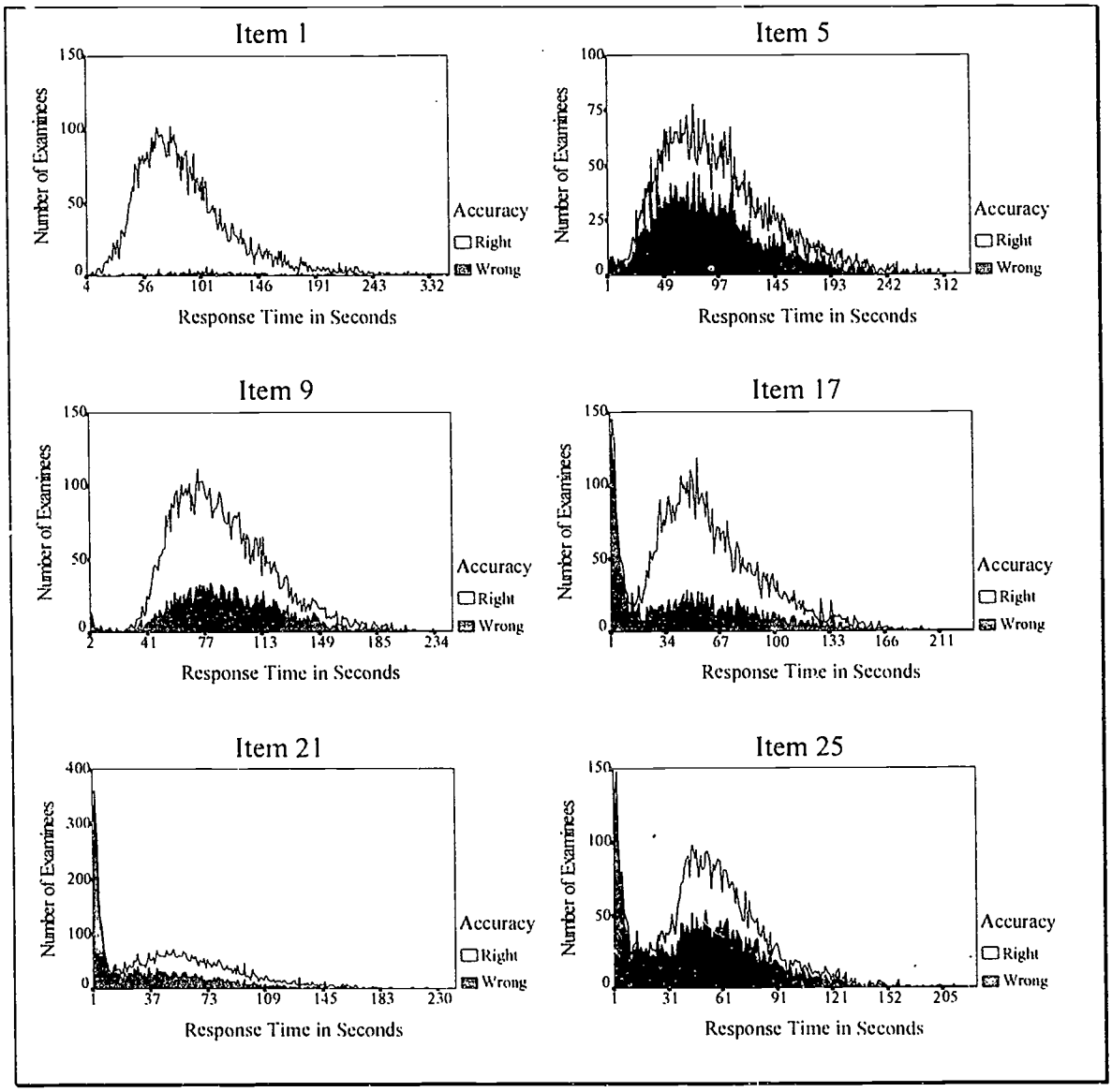

Figure 3: Response time distributions from selected items in Analytical 1. Notice the large "bump" of fast wrong responses on later items, indicating rapid-guessing behavior. (Also notice that the scales on both axes are different for each item.)

Items 17, 21, and 25 (out of 25 items; Figure 3) all show fairly high numbers of examinees engaging in rapid-guessing behavior as indicated by the large "bump" of fast wrong responses. All of the items on the last half of the test show a fairly large number of rapid guessers.

For each item, the number of examinees engaging in rapid-guessing behavior (i.e., the number of examinees in the "bump") was estimated after establishing a criterion using the response times and accuracy rates[4] to determine which responses are the result of rapid guessing (i.e., are in the "bump"). The number of examinees engaging in rapid-guessing behavior (separated by accuracy – right or wrong) and the number of examinees who did not reach the item are plotted for each item in a stacked bar chart in Figure 4. As expected, responses by examinees engaging in rapid-guessing behavior are primarily wrong responses, and in fact, accuracy is generally at or below chance (0.20 for 5 alternative multiple choice items), supporting the notion that the behavior is indeed guessing behavior.

[4] The criterion was established by determining via visual inspection where the two distributions (the rapid-guessing behavior distribution and the solution behavior distribution) crossed. Another method of determining which part of the distribution a response is in (either the rapid-guessing part or solution-behavior part) is to use a two-state mixture model (Luce, 1986; Townsend & Ashby, 1983). The criterion established via visual inspection was determined to be adequate in the present study given the purpose.

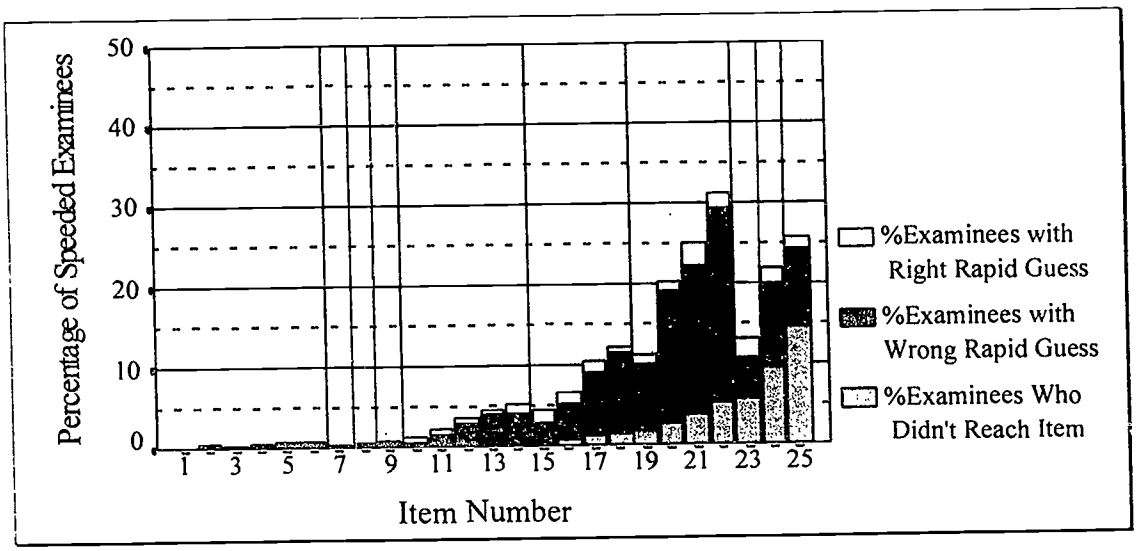

Figure 4: Percentage of speeded examinees in Analytical 1, classified into those who did not reach the item and those engaging in rapid-guessing behavior (further classified by accuracy – right or wrong). Notice that the scale ranges from 0 to 50%. Vertical lines indicate item sets. (Items 7, 8, and 9 and 23, 24, and 25 are not in sets.)

Referring back to Figure 3, the response times, in general, form a positively skewed distribution which becomes more nearly normal after the natural logarithm transformation is applied, as shown for the same items in Figure 5.[5] Rapid-guessing behavior (the "bump") creates the negative skew in the distributions of the natural logarithm of response times, ln(RT), on the items in the second half of the test. Because the ln(RT) distributions are more normal than the nontransformed response time distributions, the ln(RT)'s were used as the dependent variable in the analyses described below.

[5] Rights and wrongs are combined in Figure 5.

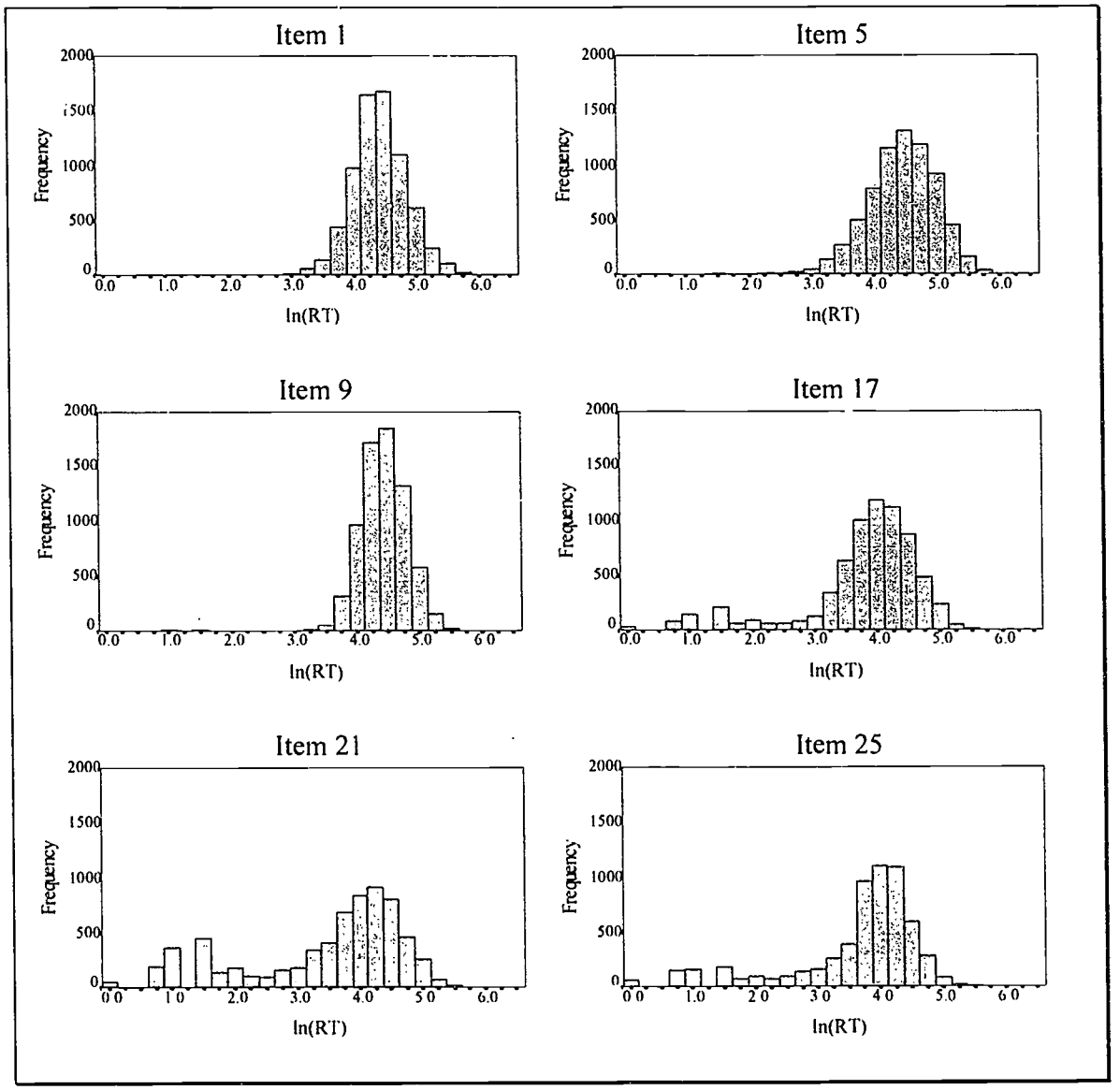

Figure 5: Natural logarithm of response times for selected items on Analytical 1. Rapid-guessing behavior creates a negative skew on later items. Otherwise, the distributions are fairly normal. (The scales on both axes are the same.)

Analyses Based on Response Times

Visual inspection of the item response time distributions supports the notion that there are two distinct behaviors occurring, solution behavior and rapid-guessing behavior. To test the notion more rigorously, a series of analyses of variance (ANOVAs) were performed and specific hypotheses were tested. In all ANOVAs, ln(RT) was predicted from various other variables, such as response and item characteristics.

Because there are so many degrees of freedom in the analyses, F-tests would be misleading (almost everything is significant). Instead, effect sizes were used to estimate the relative importance of the main effects and interactions. Partial eta squared (η_p^2) was used to estimate effect size. η_p^2 estimates the proportion of variance explained in ln(RT) by each factor, after accounting for other variables and interactions. More specifically,

$$\eta_p^2 = \frac{SSA}{SSA + SSE}$$

for Factor A, where SSA is the sum of squares for Factor A and SSE is the sum of squares for error.

The squared multiple correlation coefficient, R^2, is also provided. R^2 estimates the proportion of the total variability in ln(RT) that is explained by the model. More specifically (for a model with 2 factors, A and B),

$$R^2 = \frac{SSA + SSB + SSAB}{SSA + SSB + SSAB + SSE}$$

where SSAB is the sum of squares for the interaction between A and B and the other terms are as above.

First half of the test vs. last half of the test

The first comparison was between the first half of the test (items 1-12) and the last half of the test (items 13-25).[6] The ln(RT) was predicted from *item* (item number treated categorically) and *accuracy* (right/wrong). The *item* variable allows each item to have its own mean and standard deviation. The *item* variable is an amalgamation of all the characteristics that make each item different. Thus, no attempt was made to explain what makes items different; they were simply allowed to vary. If the *item* variable has an effect, it means that characteristics of the items affect response times. Likewise, if the *item* variable has no effect, item characteristics do no affect response times, suggesting that examinees are not fully considering the items.

On speeded items, there will be many fast wrong responses. Thus, on speeded items, a wrong response would be more likely to be fast. Therefore, the effect of *accuracy* should be greater on speeded items than on nonspeeded items.

Taken together, a decrease in the effect of *item* and an increase in the effect of *accuracy* suggest speededness. Table 2 shows the results of separate ANOVAs for the first half of the test (items 1-12) and the last half of the test (items 13-25). The last half of the test has a smaller effect of *item* and a larger effect of *accuracy* than the first half of the test. R^2 is also smaller in the last half of the test which means that, overall, response times were not explained as well on the last half of the test by item characteristics and

---

[6] The items were divided after item 12 because that divided the test as closely as possible into halves. Also, before item 12, there was a very small amount of rapid-guessing behavior, and afterwards there was an increasing amount of rapid-guessing behavior (i.e., speededness).

accuracy (right/wrong). The results suggest that the second half of the test is more speeded than the first half of the test.

Table 2: ANOVA results comparing first half of test to last half of test. The decrease in effect size for *item* and increase in effect size for *accuracy* from the first half to the last half of the test indicate that the last half of the test is somewhat speeded. The decrease in R^2 also indicates this.

	R^2	Effect Size (η_p^2)		
		Item	Accuracy	Item by Accuracy
Items 1-12	.348	.211	.002	.033
Items 13-25	.242	.148	.068	.019

The mean ln(RT) is plotted across items by accuracy (right/wrong) in Figure 6. During the first half of the test, right and wrong responses were made at about the same rate. However, on the items that had examinees engaging in rapid-guessing behavior, the wrong responses are faster, on average. In general, the larger the number of examinees engaging in rapid-guessing behavior, the bigger the difference between right and wrong response times.

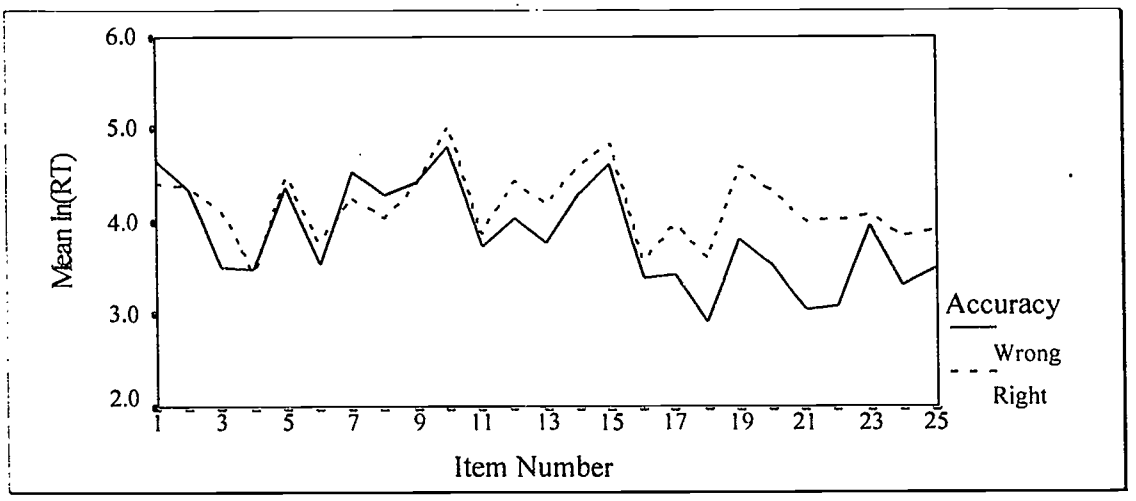

Figure 6: Mean ln(RT) across items by accuracy. Notice that wrong responses are faster than right responses starting with item 12. The difference is more pronounced after item 15, which corresponds to the number of people engaging in rapid-guessing behavior.

Without responses due to rapid-guessing behavior

Removing item responses that appeared to be due to rapid-guessing behavior from the data set should cause R^2 and the effect of *item* to increase and the effect of *accuracy* to decrease, as compared to the analyses that included responses due to rapid-guessing behavior. The difference should be more pronounced on items 13-25 where the test was more speeded. Table 3 and Table 4 show ANOVA results from the first half of the test (items 1-12) and the last half (items 13-25), respectively. After removing responses due to rapid-guessing behavior, R^2 and the effect of *item* increased and the effect of *accuracy* decreased on both halves of the test. The changes were more dramatic on the last half of the test, as predicted.

Table 3: ANOVA results for the first half of the test (items 1-12) for the complete sample and after removing the speeded responses (those resulting from rapid-guessing behavior). There are small increases in R^2 and the effect size of *item* and a small decrease in the effect size of *accuracy*, suggesting that there was a small amount of speededness on the first half of the test.

Items 1-12	R^2	Effect Size (η_p^2)		
		Item	Accuracy	Item by Accuracy
Complete Sample	.348	.211	.002	.033
Speeded Responses Removed	.401	.257	.000	.041

Table 4: ANOVA results for the last half of the test (items 13-25) for the complete sample and after removing the speeded responses (those resulting from rapid-guessing behavior). There are increases in F^2 and the effect size of *item* and a decrease in the effect size of *accuracy*, suggesting that the last half of the test was fairly speeded.

Items 13-25	R^2	Effect Size (η_p^2)		
		Item	Accuracy	Item by Accuracy
Complete Sample	.242	.148	.068	.019
Speeded Responses Removed	.336	.283	.010	.014

The mean ln(RT) is plotted across items by accuracy (right/wrong) in Figure 7 without the responses due to rapid-guessing behavior. Removing these speeded responses made the difference in ln(RT) between right and wrong responses much less pronounced, supporting the ANOVA results.

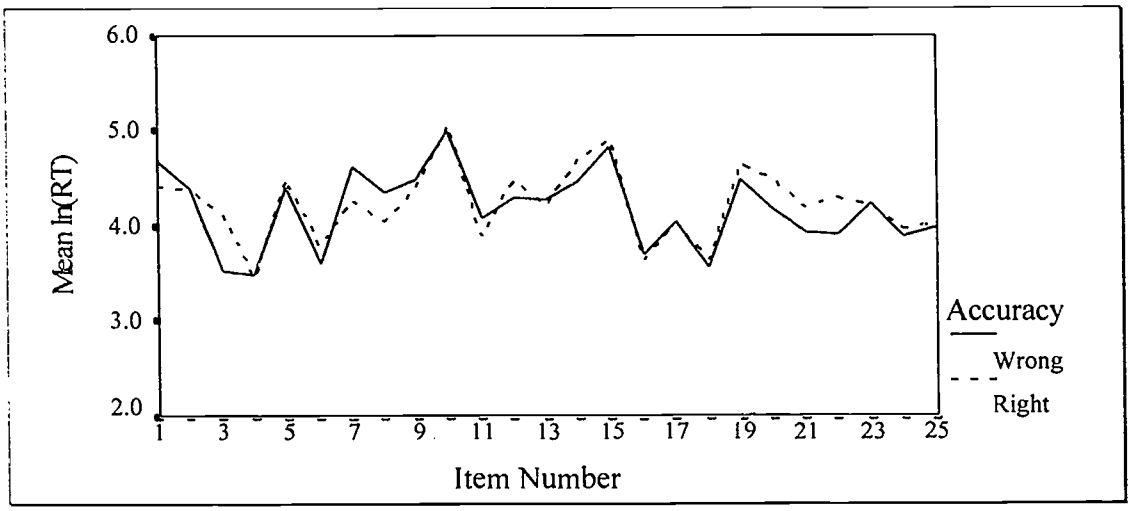

Figure 7: Mean ln(RT) across items by accuracy, after responses resulting from rapid-guessing behavior have been removed, item by item. The wrong responses are no longer faster than right responses from item 12 to item 19. The wrong responses are a little faster than right responses after item 19, but the difference is much less pronounced than when speeded responses were included.

Speeded responses marked and used to predict ln(RT)

Another way to look at the consequences of speeded responses is to mark all of the responses as speeded or not and include the new dummy-coded variable, *speeded*, in the ANOVA to explain ln(RT). Because the *speeded* variable indicates whether a response was very fast or not, the variable will obviously[7] increase R^2 when explaining Ln(RT). Therefore, an arbitrary cutoff (also dummy coded) was used to serve as a baseline against which the *speeded* cutoff results could be compared. The mean ln(RT) across all items was used as the arbitrary cutoff for each item (responses were coded as above or below the mean). The *speeded* cutoff was determined separately for each item

[7] The *speeded* variable can be thought of as a dichotomous response time. Unspeeded responses are, by definition, slower than speeded responses, thus the *speeded* variable should be an effective predictor of response time.

and was the same cutoff that was used in the previous analyses to classify responses as speeded or not.

Table 5 shows ANOVA results from the first half of the test (items 1-12) for the *speeded* cutoff and the arbitrary cutoff (both with and without rapid guesses). In the first row, responses were dummy coded as rapid guesses or not (i.e., speeded or not) and added to the ANOVA model with *item* and *accuracy*. R^2 increased slightly, as compared to the model without the *speeded* cutoff (first row of Table 2), but the increase is not very large. Similarly, the effect size of *speeded* (labeled *cutoff* in the table) was very small. This is not surprising because on these items (1-12), the *speeded* cutoff identified only 1% of the responses.

In the second row of Table 5, results are shown for the arbitrary cutoff (mean ln(RT) across all items, 4.15 ln seconds). R^2 is much larger when using the arbitrary cutoff than when using the *speeded* cutoff, and the effect size of the arbitrary cutoff is substantially larger than the effect size for the *speeded* cutoff. The arbitrary cutoff divided the response time distributions more evenly – 56% of the responses were classified as above the mean – thus it is not surprising that the arbitrary cutoff was so effective in explaining ln(RT).

Finally, the speeded responses (those arising from rapid-guessing behavior) were removed from the data set and the ANOVA with the arbitrary cutoff (4.15 ln seconds) was performed again. The results are shown in the last row in Table 5. R^2 increased a little, compared to the analysis with the speeded responses included. The effect of the arbitrary cutoff decreased slightly after removing speeded responses, but the effect is still large.

Overall, the *speeded* cutoff does not predict response times very well on the first half of the test (items 1-12), but this is not surprising because there are so few speeded responses (only 1% of the responses result from rapid guessing). The arbitrary cutoff, which divides the ln(RT) distributions nearly in half, predicts ln(RT) quite well, and this is not surprising because the arbitrary cutoff provides strong information about ln(RT) (i.e., whether it was above or below the mean).

Table 5: ANOVA results for the first half of the test (items 1-12) for three conditions: in the first row, *cutoff* indicates whether a response was a rapid guess or not; in the second row, *cutoff* indicates whether a response was above or below the mean ln(RT) across all items; in the last row, responses resulting from rapid guessing were removed, and *cutoff* again indicates whether a response was above or below the mean ln(RT) across all items.

Items 1-12	R^2	Effect Size (η_p^2)						
		Item	Accuracy	Cutoff	Item by Accuracy	Item by Cutoff	Accuracy by Cutoff	Item by Accuracy by Cutoff
Speeded response cutoff	.474	.014	.000	.019	.029	.001	.000	.000
Arbitrary cutoff	.687	.039	.005	.337	.008	.035	.014	.008
Arbitrary cutoff; no speeded responses	.726	.062	.000	.290	.010	.034	.005	.010

Results are very different on the last half of the test (items 13-25), as shown in Table 6. On the last half of the test, the *speeded* cutoff predicts response times very well. R^2 is very large, and the effect of the *speeded* cutoff is also very large. This would not be

expected because the *speeded* cutoff does not split the distribution evenly; 16% of the responses are identified as speeded.

In the second row of the table, the results for the arbitrary cutoff are shown. Although the arbitrary cutoff splits the item distributions more evenly (45% of the responses are identified as above the mean on items 13-25), R^2 and the effect of the arbitrary cutoff are not as large as when the *speeded* cutoff is used. Thus, thinking of the total item response time distribution as a mixture of two distributions, one comprised of rapid-guessing behavior and the other of solution behavior, knowing which of the two distributions a response is in provides better information about ln(RT) than knowing whether the response was above or below the overall mean ln(RT).

If the responses due to rapid guessing are removed (last row of Table 6), R^2 and the effect of the arbitrary cutoff increase as compared to when the speeded responses are included, although they are still not quite as large as when the *speeded* cutoff is used.

Clearly, there is something special about the *speeded* cutoff. The *speeded* cutoff divides the total response time distribution for each item into two logical, distinct distributions, namely the distribution resulting from rapid-guessing behavior and the distribution resulting from solution behavior. The two distributions have their own means and standard deviations, and the speeded cutoff capitalizes on this.

Table 6: ANOVA results for the last half of the test (items 13-25) for three conditions: in the first row, *cutoff* indicates whether a response was a rapid guess or not; in the second row, *cutoff* indicates whether a response was above or below the mean ln(RT) across all items; in the last row, responses resulting from rapid guessing were removed, and *cutoff* again indicates whether a response was above or below the mean ln(RT) across all items.

Items 13-25	R^2	Effect Size ($\eta_p 2$)						
		Item	Accuracy	Cutoff	Item by Accuracy	Item by Cutoff	Accuracy by Cutoff	Item by Accuracy by Cutoff
Speeded response cutoff	.754	.057	.001	.505	.001	.006	.000	.001
Arbitrary cutoff	.582	.012	.028	.393	.005	.014	.029	.007
Arbitrary cutoff; no speeded responses	.704	.054	.005	.484	.002	.018	.005	.002

SUMMARY AND CONCLUSIONS

When examinees do not have enough time to fully consider and answer all of the items on a test, they will either not finish all the items or rapidly guess on the remaining items. Both behaviors are the result of speededness, but traditional measures of speededness have taken into account only the first behavior, not finishing. In traditional, paper-and-pencil testing, it is not possible to identify responses that result from rapid-guessing behavior because such behavior can only be identified by considering both accuracy and response time. Because it is not possible to collect response time data in operational paper-and-pencil tests, using the number of unreached items to measure speededness has been the best approach available.

On computer-administered tests, it is possible to collect item response times. The purpose of the present study was to see if response times could be used to improve the assessment of speededness. The present analyses suggest that response time can be an effective tool for measuring speededness at the item level. The analyses of an Analytical section from the GRE-CBT suggest that rapid guessing was more prominent on the second half of the test than on the first half. The variation in items was less predictive of response times on the second half of the test, implying that some examinees were not fully considering the items (i.e., they were engaging in rapid-guessing behavior). Removing the responses that were due to rapid-guessing behavior caused a considerable increase in the ability of the variation between items to predict response times on the last half of the test, indicating that the remaining responses resulted from solution behavior. Finally, identifying responses as speeded was a better predictor of response times than was an arbitrary cutoff that split the response time distributions more nearly in half, suggesting that the rapid-guessing behavior is very different from solution behavior. Figure 4, which plotted across items the number of examinees who did not reach the item and the number of examinees who engaged in rapid-guessing behavior on the item, showed one way of displaying the total amount of speededness in the test at the item level.

One of the primary purposes of testing is to measure ability as accurately as possible. When speeded behavior goes undetected, we allow a time factor to contaminate the ability estimates. Now that we have the ability to detect rapid-guessing behavior, we can use this information to remove speeded responses for examinees when estimating

their abilities, as Lord (1980) suggested, and doing so should provide a more accurate estimate of ability.

A related concern is the effect of rapid guessing on item parameters. Both classical test theory and item response theory assume that all of the examinees fully considered every item. An incorrect answer is taken to mean that the examinee was unable to answer the item (i.e., the item was too difficult for the examinee). However, if a test is speeded, an incorrect answer may mean that the examinee did not have time to fully consider the item; the examinee may have been fully capable of answering the item correctly, given more time. Oshima (1994) has shown that both not answering items and randomly guessing on items causes the item discrimination and difficulty parameters to be overestimated and the guessing parameter to be underestimated. It is clear that when the assumption that examinees are fully considering each item is not met (i.e., the test is speeded), item parameters will be estimated erroneously for the speeded items if the speeded responses are not removed during item estimation.

Speededness is an important issue with which test developers must be concerned because speededness can cause item parameters and examinee ability to be incorrectly estimated. Identifying speeded behavior is the first step in dealing with speededness effectively. There are two ways that examinees can deal with time constraints, and these are to not finish all items or to rapidly guess on remaining items. Although both behaviors have been recognized as a consequence of speededness, until now there has been no way to identify rapid-guessing behavior. Thus, assessment of speededness has only considered one of these speeded behaviors – not finishing the test. The present

study shows that it is possible to identify rapid-guessing behavior using item response times, thus allowing a more accurate and rigorous assessment of the total amount of speededness in the test.

REFERENCES

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

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.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

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

Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement*, *31*(3), 200-219.

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: Cambridge University.