EFFECTS OF RESPONSE LATENCY ON TEST ITEM PARAMETERS USING DIFFERENT SCORING PROCEDURES

By

ADEDIWURA A. Adeyemi
yemtoy20002000@gmail.com
Department of Educational Foundation and Counselling
Faculty of Education
Obafemi Awolowo University, Ile-Ife

Abstract

The study determined the prevalence of response latency behavior among undergraduates and examined its effect on item difficulty and discrimination indices 50 items EDU311 e-exam across four scoring procedures as well as the difference in proportion of students that passed the test across scoring procedures. These were with the aim of profitably considering and integrating response latency into scoring process. The study adopted ex-post facto research design. The study population comprised part three students of a university of education in Nigeria. The sample consisted of 1200 students that were selected using stratified sampling technique with college serving as strata. The research instrument consisted of dichotomously-scored 50 items computer-based multiple choice EDU311 e-exams. Collected data were analyzed using Confidence Interval, Frequency count, Percentages, Descriptive and ANOVA statistics. The results showed that that there is prevalence of response latency among undergraduate students in the studied university of education with 25% of the students engaging rapid guessing at least once. The result also showed that the difficulty ($\hat{p}$) and discrimination (r_{pb}) indices values across the four scoring procedures differed significantly, ($F_{(3, 196)} = 11.25$, $p < 0.05$) and ($F_{(3, 196)} = 8.57$, $p < 0.05$) respectively. Furthermore the results showed that the confidence intervals for the percentages of students who succeeded in the e-exams were not significantly different from each other at the confidence level 95%.

Introduction

Assessment of student learning is an important part of any educational process. The effectiveness of instruction as well as the instructional decisions largely relies on the ability of teachers to construct and select tests and assessments that would bring about valid, reliable, and fair measures of learning outcomes (Linn & Gronlund, 1995). The nature and the quality of gathered information can control the educational development efforts and direct the instruction. Assessment of student learning in any educational settings is mainly aimed at measuring students' achievement as well as making series of decisions based on students' performance. The accuracy of such decisions may depend on the educational consequences to the students, the school, the society, policy makers, stakeholders or even the parents/guardian.

Response latency (Rapid-guessing) behavior is often observed when examinees do not put forth sufficient effort and time to answer test questions. When examinees are less motivated, they are unlikely to give adequate effort throughout the test which creates problems in presenting what they know and may result in biased estimates for ability and item parameter estimates. If a student does not put forth an adequate required effort, assessment results may underestimate the proficiency level and thus, it may lead to invalid interpretations of the obtained scores. Education policymakers and professionals often desire to use tests for multiple purposes, such as monitoring the educational system, aiding instructional planning, motivating students to perform better, acting as a mechanism to change instructional content, and holding schools and educators accountable. In addition, they use tests for certifying students as having attained specific levels of achievement (Hamilton, Stecher, & Klein, 2002). Making inferences about a student's performance goes beyond the specific test that is used. Sometimes, teachers would like to know the degree of a student's understanding of specific concepts based on the score that he or she obtains on achievement tests, which corresponds to the knowledge and skills the student learns in the usual school subjects. Achievement tests have commonly been used to measure students' educational progress, but the number of purposes they expect to serve has grown substantially. Research reports overtime had made efforts toward explaining the inconsistency between student scores and the actual student's knowledge of the subject matter. Measurement experts intend to obtain valid test scores from each test administration.

Whenever the examinee perceives that his or her performance has little or no consequences, it is more likely that he or she will put little effort into the assessment. Consequently, their scores may not accurately reflect their true abilities (Wise & DeMars, 2003, 2005). This case represents a direct threat to test score validity, and experts have a responsibility for taking corrective action.

Technology growth and recent developments in computer usage in the practice of assessment is providing great opportunities in exploring new ways that are geared towards improving the quality of assessment data (Klein & Hamilton, 1999). Collection of additional information relating to the interaction between individual examinee and each of the items that made up the test is relatively made easy with Computer Administered Tests CAT. The time required for a response can give us a good indicator of the student's effort while testing in CAT. This is referred to as the "response latency." Low-effort responses can be rescored in different ways to affect items' and persons' characteristics which may have an impact on educational decisions (Hadadi & Luecht 1998).

The use of computer to administer tests to a large extent expands the range of performance tasks that can be included in standardized tests and the information gathered during the testing session. The use of computer in test administration can easily provide additional data at the item level called "response latency," and behaviors related to the answer can be recorded. Response latency is the time used by an examinee when responding to an item. The capability to measure response latency at the item level appears to be easy and valuable. These response latencies can be used for both assessing the degree to which examinees give effort to the test and assessing the degree to which items receive good effort (Wise, 2006).

Under a typical testing setting, some examinees will not reach the last questions on a test because of the time limit. Items of this sort are referred to as not-reached. Even when items are reached and read, some examinees may appraise the item's content and decide for their own reasons not to respond. Items of this sort are referred to as omitted items. On the other hand, if the examinee realizes that there are no direct personal consequences based on his or her performance, the examinee may engage in random-guessing response behavior (Mislevy and Wu, 1996). Wise (2006) shows that there are three primary choices of effort measure: examinee self-reports, person-fit statistics, and response-time based measures.

The use of self-report instruments is the first method to detect examinee efforts, which seeks answer to questions such as, "I am not concerned about the score I receive on this test."

Administering and scoring self-report instruments are easy; however, there may be problems when inferences are drawn based on student responses. For example examinees may be giving socially acceptable responses that do not match with what they actually believe. The information gathered through self-reports may not correspond to the student's actual behavior during the examination. Also younger children may not understand the self-report questions and be incapable of giving accurate responses. In some cases, students who did not try to do well on the test might falsely report giving good effort because of fear of disapproval or punishment from the test giver. Other students who believe they did not perform well on the test might underreport their effort because of a predisposition to attribute failure on the test to lack of effort over lack of ability (Pintrich & Schunk, 2002; Wise, 2006). And, of course, students who choose not to expend the necessary energy on the test for a valid score may also choose not to expend the necessary energy on the self-report instrument for a valid score.

Examination of person-fit statistics is the second method that can be used in detecting examinee-effort. This method compares the examinee's response pattern with a theoretical measurement model. This method has a clear advantage of being based on an observed item response pattern rather than self-report questions. The use of person-fit indices, as an exploratory technique, applied typically in situations in which the kind of aberrant responses can be expected, is unknown, leads to numerous interpretations. According to Wise and DeMars, (2006), concluding unambiguously that a particular instance of misfit is due to lack of effort may be difficult. Misfit may indicate that the examinee engaged in cheating, creative responses, careless responses, guessing, or other unusual response behaviors. This is because it cannot always be confident of the kind of aberrant responses underlying test performance because different forms of aberrant behavior can have multiple explanations and may result in the same kind of item response pattern (Meijer, 1996). Thus, it becomes difficult to rely solely on person-fit statistics to measure examinee effort.

As a result of the short comings of the first and second methods of detecting examinee effort, Wise and Kong (2005) developed an index named *Response Time Effort* (RTE) to measure the examinee's test-taking effort in computer-based settings. The developed RTE was based on series prior research reports (Schnipke, 1999 and Schnipke & Scrams 1997, 2002) that differentiates; appropriate solution-oriented behavior (where the examinees actively seek to determine the correct answer to test items) from rapid-guessing behavior (where the examinees determine the answer by rapid-guessing; perhaps because they do not have enough time to fully think about the item). Wise and Kong in their study reported that rapid-guessing behavior appears to be obvious in untimed data in low-stakes computer-based tests from the beginning to the end of the tests. This assertion is in contrast to Schnipke and Scrams earlier proposal that rapid-guessing behavior is only observed in examinee towards the end of speeded tests.

Information on examinee efforts gathered using response latency is considered more important and useful than what is revealed using self-reports. This is because the information represents a direct observation of examinee behavior and it does not rely on examinee judgments. Moreover, data for response latency are collected in an easy and non-reactive way. Examinees may have little awareness that response time data are even being recorded. Response latency information is at the item level and therefore allows tracking the changes in the level of effort during a testing session. Unlike person-fit statistics, response latency is a direct way to examine the effort that the examinee puts forth on test items and may occur along with other aberrant behaviors during the testing session. A good information and understanding of the nature of rapid-guessing behavior and the development of indices to detect this behavior can help in the process of test construction, testing of response validity, and the verification of decisions made using these test results.

This study therefore, is aimed at detecting occurrence of a type of guessing that is expected to occur if the examinee does not put forth adequate effort as well as responding to an item without reading it using response latency. Three different scoring methods were used to replace the guessed responses instead of deleting them, and attempts were made to determine how these scoring procedures impact different parameter estimations for items in terms of CTT models. Items on which examinees spend less than a predetermined threshold of time were flagged and rescored in a variety of ways including: not reached, omitted, zero, or the response remained as the examinees answered it, and scored 1 if correct, otherwise 0. There after the rescored data sets were used to determine if these different rescoring procedures effectively change the relevant classical test theory (CTT) parameter estimates and decisions based on examinees' scores compared with the original data. The study objective is to evaluate the use of different scoring procedures on CTT parameter estimates for dichotomously scored items that are obtained from computer-administration after identifying rapid-guessing responses.

Objectives

The specific objectives of the study were to:

1. determine the prevalence of rapid guessing (response latency) behavior among university undergraduates;
2. examine the effect of response latency on item difficulty of dichotomously-scored 50 items computer-based multiple choice EDU311 e-exam across different scoring procedures
3. establish the effect of response latency on item discrimination across different scoring procedures; and
4. to determine the difference in proportion of students that passed the test across scoring procedures

Research Question: What is the prevalence of response latency among secondary school Students?

Hypotheses

1. There is no significant difference in the item difficulty indices of dichotomously-scored 50 items computer-based multiple choice EDU311 e-examination across scoring procedures

2. There is no significant difference in the item discrimination indices of dichotomously-scored 50 items computer-based multiple choice EDU311 e-examination across scoring procedures

3. There is no significant difference in proportion of students that passed the test across scoring procedures

Method

The study adopted ex-post facto research design in which data that accrued from students' responses in an education course e-examination of a university of education were used without contact with the students involved. The study population comprised part three students of a university of education in Nigeria. The sample consisted of 1200 students that were selected using stratified sampling technique with college serving as strata. From each of the four colleges of the university 300 part three students that registered for an Education general course EDU 311 (Introduction to Educational Research Method) were randomly selected. The research instrument consisted of dichotomously-scored 50 items computer-based multiple choice EDU311 e-exams. Individual item responses and the amount of time taken by the examinees to read, review, and answer individual items were recorded for each student. Examinee's response and response time latency for each item on the test were recorded. The software recorded the response time latency as the number of seconds elapsed between the display of the item on the screen and an examinee's submission of the response. Items that were answered in less time than the latency threshold were identified and responses were rescored using one of the suggested scoring procedures. Collected data were then analyzed to determine whether the rescored data results in different parameter estimates for items. A classical item analysis was conducted to find item difficulty estimates, percentage of students selecting the correct answer, and item discrimination indexes. Point-biserial correlation coefficients were computed for each test item. The differences in item parameter estimates: difficulty, and discrimination, calibrated with regard to different scoring methods were examined utilizing simple ANOVA statistics. Post-hoc analysis was employed to facilitate comparisons and to find the significance of differences, if any, between parameter estimates for items across different scoring procedures.

Results

Research Question 1: What is the prevalence of response latency among undergraduate students?

To answer this question, the confidence interval (CI) for the proportion of guessing committed by students on each of the item was constructed using the following formula;

The CI for a proportion,

$$\pi = P - [Z_{1-\alpha/2} * SE_p] \text{ to } P + [Z_{1-\alpha/2} * SE_p]$$

Where P is the proportion of examinees committed rapid-guessing behavior,

$Z_{1-\alpha/2}$ is the percentile from the standard normal distribution. Thus, for a 95% CI $Z_{1-\alpha/2} = 1.96$. SE_p, the standard error of the proportion, is equal to

$$SEp = \sqrt{\frac{p(1-p)}{N}}$$

Where N is the number of observations.

The result is as presented in Table 1

Table 1: Confidence intervals for proportion of guessing committed based on item level

Item No.	Guessing (*p*) %	*Se*p	L95%*	U95%**
1	0.099	0.012	0.075	0.123
2	0.085	0.012	0.063	0.108
3	0.029	0.007	0.015	0.043
4	0.039	0.008	0.024	0.055
5	0.049	0.009	0.032	0.067
6	0.020	0.006	0.009	0.032
7	0.036	0.008	0.021	0.051
8	0.031	0.007	0.017	0.045
9	0.099	0.012	0.075	0.123
10	0.070	0.011	0.049	0.091
11	0.038	0.008	0.022	0.053
12	0.048	0.009	0.031	0.065
13	0.354	0.039	0.256	0.352

14	0.038	0.008	0.022	0.053
15	0.080	0.011	0.058	0.102
16	0.345	0.026	0.299	0.381
17	0.055	0.009	0.036	0.073
18	0.048	0.009	0.031	0.065
19	0.130	0.014	0.102	0.157
20	0.073	0.011	0.052	0.094
21	0.128	0.014	0.101	0.155
22	0.302	0.019	0.265	0.339
23	0.121	0.013	0.095	0.148
24	0.364	0.039	0.276	0.365
25	0.111	0.013	0.086	0.136
26	0.177	0.016	0.147	0.208
27	0.147	0.015	0.118	0.175
28	0.253	0.018	0.217	0.288
29	0.131	0.014	0.104	0.159
30	0.130	0.014	0.102	0.157
31	0.049	0.009	0.032	0.067
32	0.344	0.021	0.296	0.374
33	0.036	0.008	0.021	0.051
34	0.031	0.007	0.017	0.045
35	0.364	0.019	0.286	0.362
36	0.206	0.027	0.194	0.269
37	0.162	0.015	0.132	0.192
38	0.048	0.009	0.031	0.065
39	0.114	0.013	0.089	0.140
40	0.038	0.008	0.022	0.053
41	0.130	0.014	0.102	0.157
42	0.324	0.019	0.286	0.362
43	0.020	0.006	0.009	0.032
44	0.036	0.008	0.021	0.051
45	0.031	0.007	0.017	0.045
46	0.225	0.017	0.191	0.259
47	0.169	0.015	0.139	0.199
48	0.164	0.015	0.134	0.194
49	0.324	0.019	0.286	0.362
50	0.177	0.016	0.147	0.208

Table 1 showed the proportion of guessing among the undergraduates which indicated that rapid-guessing behavior was present in all items since none of the CIs contains the value 0. However, rapid guessing responses were committed more on items 13, 16, 24, 32, 36, 42 and 49 of the test and the lower limits of the confidence intervals turn out to be far from capturing the null value of zero. The result as presented in Table 1 is an indication that there is prevalence of response latency among undergraduate students in the studied university of education. The prevalence of response latency was further established through a further analysis of number of rapid-guessing responses across Examinees. The result is as presented in Table 2

Table 2: Frequency of examinees exhibited rapid-guessing behavior

Number of rapid guesses	Frequency	Percent
1 – 10	300	25
11 – 20	251	20.9
21 – 30	283	23.6
31- 40	173	14.4
41 – 50	193	16.1
Total	1200	100.0

The result as presented in Table 2 showed that there is none out of the sampled 1200 students that did not engaged rapid guessing in their attempt to provide answers to each of the 50 items. The Table showed that 25% of the students engaged rapid guessing at least once and at most 10 times. It could also be observed that 16.1% of the students engaged rapid guessing on 41 to 50 items of the test. Thus it could be concluded that there is the prevalence of response latency among the undergraduate students in their responses to dichotomously-scored 50 items computer-based multiple choice EDU311 e-exam.

Hypothesis 1: There is no significant difference in the item difficulty indices of dichotomously-scored 50 items multiple choice EDU311 e-examinations across scoring procedures

To test this hypothesis, a classical item analysis to determine the item difficulty (*p)* of each of the 50 items was carried out across the scoring procedures (Original response, Omitted, Not presented and Zero). Table 3 presented the descriptive statistics of the 50 multiple choice EDU 311 difficulty indices across the four scoring procedures.

Table 3: Descriptive statistics of the 50 multiple choice EDU 311 difficulty indices

Scoring procedure	N	Min	Max	$\bar{x}$	SD
Original response	50	0.23	0.84	0.758	0.067
Zero	50	0.24	0.84	0.601	0.147
Not-presented	50	0.21	0.84	0.759	0.067
Omitted	50	0.35	0.91	.0.808	0.055

The item analysis result as presented in Table 3 showed that item difficulty $\hat{p}$, for Original-response ranged between 0.23 to 0.84 with a mean and standard deviation value ($\bar{x}$ = 0.758, SD = 0.067). A close observation of the means of $\hat{p}$ values for the other scoring procedure as presented in Table 3 showed that the means of $\hat{p}$ values varies from that of the original scoring procedure except for the “Not-presented” procedure where the mean of $\hat{p}$ value is similar to the “Original response” scoring procedure. The strength of the variation is then determined with the use of One-Way Analysis of Variance (ANOVA) statistic. The result of the ANOVA statistic is as presented in Table 4.

Table 4: ANOVA statistics summary showing the difference in mean of difficulty indices

Source of Variation	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.286	3	.095	11.245	.000
Within Groups	1.662	196	.008		
Total	1.948	199			

The ANOVA statistical analysis results as presented in Table 4 showed that the difficulty indices (pˆ) values across the four scoring procedures differed significantly, ($F_{(3, 196)} = 11.25$, $p <$ 0.05). Using the "Original response" scoring procedure as reference measure, since the aim of the study was to compare the original response scoring procedure with the other three scoring procedures, a multiple comparison was used to compare difficulty indices (pˆ) value estimates obtained from the scoring procedures. The result is as presented in Table 5

Table 5: Post-hoc multiple comparison difficulty indices estimates across scoring procedures

(I) Scoring Procedures	(J) Scoring Procedures	Mean Difference (I-J)	Std. Error	Sig.
Original	Zero	.0570*	.01842	.002
	Omitted	-.0498*	.01842	.007
	Not Presented	-.0010	.01842	.957

The multiple comparison result as presented in Table 5 showed that the mean differences (I-J) in difficulty estimates between the original scoring procedure and the "Zero "as well as "Omitted" procedures were significant (I-J = 0.057 and -0.0498, p<0.05) respectively. However, the mean difference (I-J) in difficulty estimates between the original scoring procedure and the Not-presented procedure is not significant (I-J = -0.0010, p > 0.05).

Hypothesis 2: There is no significant difference in the item discrimination indices of dichotomously-scored 50 items multiple choice EDU311 e-examinations across scoring procedures

To test this hypothesis, item analysis was carried out on each of the 50 items across the scoring procedures (Original response, Omitted, Not presented and Zero) to estimate the discrimination indices (r_{pb}). Table 6 presented the descriptive statistics of the 50 multiple choice EDU 311 discrimination indices (r_{pb}) across the four scoring procedures.

Table 6: Descriptive statistics of the 50 multiple choice EDU 311 discrimination indices

Scoring procedure	N	Min	Max	$\bar{x}$	SD
Original response	50	-.23	.77	0.292	0.198
Zero	50	-.23	.77	0.167	0.151
Omitted	50	-.23	.77	0.417	0.192
Not-presented	50	-.23	.77	0.295	0.181

The item analysis result as presented in Table 6 showed that item discrimination (r_{pb}) indices for all the scoring procedures ranged from -0.23 to 0.77. However the mean and standard deviation for the four scoring procedure defer. While the mean and standard deviation values of r_{pb} for the "Original response" and "Not-presented" were respectively ($\bar{x}$ = 0.292, SD = 0.198) and ($\bar{x}$ = 0.292, SD = 0.198), the mean and standard deviation values of r_{pb} for the "Zero" and "Omitted" scoring procedure respectively were ($\bar{x}$ = 0.238, SD = 0.151) and ($\bar{x}$ = 0.417, SD = 0.192). The strength of the difference in the means of r_{pb} were then determined with the use of One-Way Analysis of Variance (ANOVA) statistic. The result of the ANOVA statistic is as presented in Table 7.

Table 7: ANOVA statistics summary showing the difference in mean of discrimination indices estimates

Source of Variation	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.847	3	.282	8.573	.000
Within Groups	6.455	196	.033		
Total	7.302	199			

.

The ANOVA statistical analysis results as presented in Table 7 showed that the discrimination indices (r_{pb}) values across the four scoring procedures differed significantly, ($F_{(3, 196)} = 8.57$, $p < 0.05$). Using the "Original response" scoring procedure as reference measure, Scheffe multiple comparison Post-hoc analysis was used to compare the discrimination indices (r_{pb}) value estimates obtained from the scoring procedures. The result is as presented in Table 8

Table 8: Post-hoc multiple comparison discrimination indices estimates across scoring procedures

(I) Scoring Procedures	(J) Scoring Procedures	Mean Difference (I-J)	Std. Error	Sig.
Original	Zero	.12430*	.03710	.013
	Omitted	-.12500*	.03710	.011
	Not Presented	-.01000	.03710	.995

The multiple comparison result as presented in Table 8 showed that the mean differences (I-J) in discrimination estimates between the original scoring procedure and the "Zero "as well as "Omitted" procedures were significant (I-J = 0.124 and -0.125, $p < 0.05$) respectively. However, the mean difference (I-J) in discrimination estimates between the original scoring procedure and the Not-presented procedure was not significant (I-J = -0.010, $p > 0.05$).

Hypothesis 3: There is no significant difference in proportion of students that passed the test across scoring procedures.

To test this hypothesis, three arbitrary cut-scores (60%, 70% and 80%) were specified and proportion of students that passed the test were computed based on the three arbitrary cut-scores. The result was as presented in Table 9.

Table 9: Proportion of students that passed the test across scoring procedure based on different cut-scores

Scoring procedure	**Cut-score**		
	60%	**70%**	**80%**
Original response	42.9	28.6	13.3
Zero	41.5	27.4	12.7
Omitted	39.3	26.8	12.4
Not-presented	51.5	27.3	14.4

Table 9 presented the proportion of students that pass the dichotomously-scored 50 items multiple choice EDU311 e-examinations across scoring procedures, according to the three cut-scores. The result showed that the proportion of students that passed the e-exams was similar for all scoring procedures when 70% and 80% correct were used as cut-scores. However, when 60% correct was used as a cut-score the proportions of students that passed the exam were dissimilar for the Not-presented procedure. Table 9 showed that at least about 9% of the students were classified differently based on the Not-presented procedure compared with the original responses. Test of significance difference from zero was then carried out on the proportion difference of 9% using a 95% CI for the difference between paired proportions with aid Confidence Interval Analysis software (Bryant, 2000). The result was as presented in Table 10

Table 10: 95% CI for the proportion of students that passed across scoring procedures

Scoring procedure	**Cut-score**					
	60%		**70%**		**80%**	
	L95%	U95%	L95%	U95%	L95%	U95%
Original response	0.295	0.360	0.157	0.204	0.105	0.131
Zero	0.285	0.351	0.155	0.199	0.103	0.130
Omitted	0.290	0.358	0.155	0.199	0.103	0.130
Not-presented	0.330	0.402	0.175	0.223	0.114	0.140

The result as presented in Table 10 showed that the CI limits does not contain the null value of zero with 95% CI that ranged between 0.035 to 0.042. The resulting interval suggests that in the population at least 3.5% to 4.2% of the sampled students may be classified differently if we utilize the original responses rather than using the Not-presented procedure with cut-score 60% of the items. Thus, the calculated CIs for the percentages of students who succeeded in the e-exams appeared to be the same and they were not significantly different from each other at the confidence level 95%.

Discussion

The study described rapid-guessing responses among the sample and constructed confidence intervals for the proportion of examinees who exhibited response latency behavior for each dichotomously-scored 50 items multiple choice EDU311 e-examinations with aim of establishing the prevalence of response latency among the undergraduate students. The results revealed that rapid guessing is prevalent among the sampled undergraduate students. This is evidence in the findings that showed that different proportion of examinees showed rapid-guessing behavior on every of the 50 items multiple choice EDU311 e-examinations. This finding suggests a low level of effort put forth on the test items by some students. This finding was similar to earlier findings that under rapid guessing behavior the probability of a correct response did not significantly differ from what was expected and remained near the level of chance (Wise, 2006; Wise & Kong,2005).

The conduct of item analysis on each of the 50 items multiple choice EDU311 e-examinations showed that both the item difficulty and discrimination indices were significantly different for the Omitted and the Zero procedures when compared with Original-response scoring procedure. The difference as obtained in the study thus indicated the amount of rapid guessing on an item influenced the item's mean and correlation with scores of another item or the entire test. However, the difference in the difficulty and discrimination estimates between the Original scoring procedure and the Not-presented procedure was not significant. These results with respect to difficulty estimates were consistent with Wise (2006) in that rescoring rapid-guessing responses had little effect on item difficulty. In contrast, the findings of discrimination estimates were inconsistent with Wise's results.

The present study revealed that different proportions of students passed the cut-scores across the scoring procedures used in study. Some of the undergraduate students that were reported to have been unsuccessful using Original-response scoring procedure passed after rescoring the less thoughtful responses following the Not-presented procedure. The confidence intervals analysis showed at least 3.5% to 4.2% of students in the population may have been classified differently. As a result, misclassification may occur if we utilize the original responses rather than rescoring them as Not-presented. It appears that identifying individual examinee rapid-guessing responses and rescoring them may ultimately influence the scores and, therefore, the decision taken upon performance might be changed accordingly.

A possible implication of these results may be beneficial for norming and equating studies especially when conducted under low-stakes settings. It is not surprising that norms sometimes appear lower than expected because under low-stakes settings the scores underestimate the students' ability level as they fail to capture the full effort of the examinees (Wolf, Smith, & Birnbaum, 1995). The study therefore concluded that response latency need to be considered and integrated into the scoring process, because response latency differentiates between the more thoughtful responses and the raid-guessed responses.

Reference

Bryant, T. (2000). Confidence Interval Analysis (version 2.0.0) [Computer software]. University of Southampton. London, UK.

Hadadi, A. & Luecht, R. M. (1998). Some methods for detecting and understanding test speediness on timed multiple choice tests. *Academic Medicine*, *73(10)*, 47-50.

Hamilton, L., Stecher, B., & Klein, S. (2002). *Making sense of test-based accountability in education*. Pittsburgh, PA: RAND.

Klein, S. & Hamilton, L. (1999). *Large-scale testing current practices and new directions.* Pittsburgh, PA: RAND.

Linn, R., & Gronlund, N. (1995). *Measurement and assessment in teaching, 7th edition.* New Jersey: Prentice-Hall Inc.

Meijer, Rob R. (1996). Person-fit research: an introduction. *Applied Measurement in Education, 9*(1), 3-8.

Mislevy, R. J., & Wu, P. (July, 1996). Missing responses and Bayesian IRT ability estimation: Omits, choice, time limits, and adaptive testing (with P-K. Wu). (Research Report RR-96-30-NR). Princeton, NJ: Educational Testing Service.

Pintrich, P., & Schunk, D. (2002). *Motivation in education: Theory, research, and applications (2nd ed.)*. Upper Saddle River, NJ: Merrill Prentice-Hall.

Schnipke, D. L. (1999). *The influence of speededness on item-parameter estimation* (Computerized Testing Rep. No. 96–07). Princeton, NJ: Law School Admission Council. Retrieved from ERIC Documents ED467809.

Schnipke, D., & Scrams, D. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement, 34*(3), 213-232.

Schnipke, D., & Scrams, D. (2002). Exploring issues of examinee behavior: Insights gained from response-time analysis. In, C. Mills, M. Potenza, J. Fremer, & W. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 237-266). Mahwah, NJ: Lawrence Erlbaum Associates.

Wise, S. (2006). An investigation of the differential effort received by items on a low stakes computer-based test. *Applied Measurement in Education, 19*(2), 95–114.

Wise, S., & DeMars, C. (2003, June). *Examinee motivation in low-stakes assessment: Problems and potential solutions.* Paper presented at the annual meeting of the American Association of Higher Education Assessment Conference, Seattle, WA.

Wise, S., & DeMars, C. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*(1), 1–17.

Wise, S., & DeMars, C. (2006). An application of item response time: The effort moderated IRT model. *Journal of Educational Measurement, 43*(1) 19–38.

Wise, S., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(2), 163–183.

Wolf, L. F., Smith, J. K., & Birnbaum, M. E. (1995). Consequence of performance, test motivation, and mentally taxing items. *Applied Measurement in Education, 8*(4), 341-351.