ORIGINAL RESEARCH

The relationship between classical item characteristics and item response time on computer-based testing

Yoo-mi Chae[1], Seok Gun Park[1,2] and Ilyong Park[3]

Departments of [1]Medical Education, [2]Nuclear Medicine, and [3]Biomedical Engineering, Dankook University College of Medicine, Cheonan, Korea

Purpose: This study investigated the relationship between the item response time (iRT) and classic item analysis indicators obtained from computer-based test (CBT) results and deduce students' problem-solving behavior using the relationship.

Methods: We retrospectively analyzed the results of the Comprehensive Basic Medical Sciences Examination conducted for 5 years by a CBT system in Dankook University College of Medicine. iRT is defined as the time spent to answer the question. The discrimination index and the difficulty level were used to analyze the items using classical test theory (CTT). The relationship of iRT and the CTT were investigated using a correlation analysis. An analysis of variance was performed to identify the difference between iRT and difficulty level. A regression analysis was conducted to examine the effect of the difficulty index and discrimination index on iRT.

Results: iRT increases with increasing difficulty index, and iRT tends to decrease with increasing discrimination index. The students' effort is increased when they solve difficult items but reduced when they are confronted with items with a high discrimination. The students' test effort represented by iRT was properly maintained when the items have a 'desirable' difficulty and a 'good' discrimination.

Conclusion: The results of our study show that an adequate degree of item difficulty and discrimination is required to increase students' motivation. It might be inferred that with the combination of CTT and iRT, we can gain insights about the quality of the examination and test behaviors of the students, which can provide us with more powerful tools to improve them.

Key Words: Computer-based test, Difficulty index, Discrimination index, Item response time

Introduction

Paper-and-pencil based tests are widely used as the most common way to evaluate cognitive knowledge. As the validity and the reliability of the evaluation and the actualization of its purpose depends on the quality of evaluation tool [1,2], item analysis and feedback on paper-and-pencil based tests are very important to improve the quality of the assessment. Item analysis using classical test theory (CTT) is easy to understand and apply. Item analysis is a process that examines student responses to individual test items in order to assess the quality of those items and of the test as a whole. Item difficulty is relevant for determining whether students have learned the concept being tested [3].

Since 2008, the Korean Healthcare Personnel Licensing Examination Institute has been preparing to install a

Received: October 31, 2018 • Revised: December 31, 2018 • Accepted: January 9, 2019
Corresponding Author: Ilyong Park (https://orcid.org/0000-0003-1613-4209)
Department of Biomedical Engineering, Dankook University College of Medicine, 119 Dandae-ro, Dongnam-gu, Cheonan 31116, Korea
Tel: +82.41.550.1827 Fax: +82.41.551.1827 email: piyong@dankook.ac.kr

Korean J Med Educ 2019 Mar; 31(1): 1-9.
https://doi.org/10.3946/kjme.2019.113
eISSN: 2005-7288

Korean Medical Licensing Examination (KMLE) using computerized adaptive testing [4,5]. Recently, the use of the smart device-based test or the computer-based test (CBT) for KMLE has been under discussion. There are a number of advantages associated with the CBTs, such as immediate scoring and feedback, and adaptive testing [2,5]. CBT has many of the same merits as paper-and-pencil tests and, at the same time, may provide a more reality-based clinical situation from medical institutes [6]. Further, the item analysis as well as the examinee's score can be obtained immediately after testing.

Computerization allows previously unobtainable data, such as response time, to be collected and used to improve tests [7]. Response time can be used to infer the existence of examinee strategies [8]. Previous studies have found several variables to be significant predictors of examinee response time for a single item. Response time increases with increasing item text length and increasing item difficulty [7,9,10]. Response time also varies by content category, whether the item contained an illustration or a distractor position of the correct response [7,9] and whether the examinee got the item correct or not [7]. Examinee variables (test anxiety, gender, ethnic background, age, and language) accounted for an additional 2% of the variance in response time [7]. Wise and Kong [11] introduced a measure of examinee effort based on item response times (iRTs) in CBT. Wise [12] reported that the motivation levels of examinees in low-stakes CBT are often a matter of concern to test givers because a lack of examinee effort represents a direct threat to the validity of the test data. Most studies exploring the relationship between CBT item characteristics and iRT focused on the examinee's guessing behavior in the low-stake test of which results are not reflected to the grades.

In this study, the authors investigated the relationship between the problem-solving behavior of students using the iRT and the classic item analysis indicators obtained from the CBT results. To our knowledge, there has been no report using iRT in Korea. Further, the comparison between the classic item analysis results and the iRT obtained from the CBT test has also not yet been reported in Korea.

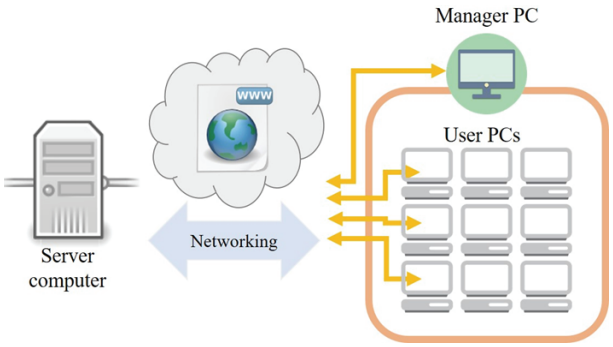
The total number of Comprehensive Basic Medical Sciences Examination (CBMSE) items assessed through CBT at Dankook University College of Medicine has roughly reached 3,500. iRT and classic item analysis data for all question items have been accumulated. In this study, we investigated the relationship between iRT and the item analysis result derived from CTT and examined the issue of examinee effort according to the item difficulty and discrimination level. We discussed the applicability of iRT in addition to the classic item analysis indicators in CBT to improve the quality of the tests. This study was approved by the Institutional Review Board of Dankook University Hospital, Korea (DKUHIRB No. 2018-10-024).

Methods

1. Context and materials

In December 2009, we implemented a CBT system that has a web-based server-and-client structure similar to the one seen in Fig. 1. The CBT system was applied to CBMSE organized successfully by Dankook University College of Medicine. Since 2010, Dankook University College of Medicine has conducted CBMSE using the evaluation test for basic medical science developed by the Medical Education Evaluation Team of the Korean Association of Medical Colleges. The evaluation test is composed of 260 items in seven courses such as

Fig. 1. Basic Configuration of the Web-Based Computer-Based Test System for This Study



physiology (35 items), biochemistry (35 items), anatomy (60 items), pathology (45 items), parasitology (15 items), pharmacology (35 items), and microbiology (35 items). In this study, the classic item analysis is performed on all items of the evaluation tests from 2013 to 2017. All of the items are multiple choice with five answer branches.

To prevent cheating, the CBT system was programmed to show the test item questions in a random order for each student. At the same time, the choices for each question were also displayed in a random order. The numbers of students for the evaluation tests from 2013 to 2017 were 44, 42, 41, 38, and 43, respectively.

2. Data source and measures

We retrospectively analyzed the results of the CBMSE conducted from 2013 to 2017 using the CBT system of Dankook University College of Medicine. The data used in the analysis are test year, course name, test date, test start and end times, number of participants, number of questions, the classic analysis results (difficulty and discrimination), average score, and the average response time taken for each item. iRT is defined as the time spent to answer the question and the difference between the time at which the student first began to solve the question and the time at which the final answer was entered. If answers to the items that have already been filled are changed during the test, iRT is re-calculated and replaced the former response time.

Among the item analysis methods using the CTT, the discrimination index of the item was defined using the two-point correlation coefficient, and the difficulty level was the percentage of the students who answered correctly out of all of the participants. The difficulty level of the item is in the range of 0-1 (the higher value, the easier question). If the difficulty level is 0-0.3, it is classified as a difficult item; 0.3-0.8 is classified as a desirable item, and over 0.8 is classified as an easy item [13]. The discrimination index ranges from 0-1; closer to 1 indicates a higher discrimination value. If it is less than 0.2, it is a poor discrimination item; 0.2-0.29 is a fair item, 0.3-0.39 is a good item, and over 0.4 is a very good item with a high degree of discrimination [13].

3. Statistical analysis

The relationship between iRT and classic item analysis was analyzed using descriptive statistics in IBM SPSS for Windows ver. 24.0 (IBM Corp., Armonk, USA). The relationship among iRT, total test time, item length, and CTT were investigated using correlation analyses. An analysis of variance was performed to identify the difference of iRT and difficulty level according to different subjects. A regression analysis was conducted to examine the effect of the difficulty index and discrimination index on iRT.

Results

1. General characteristics

The total number of questions in each of the seven courses from 2013 to 2017 ranged from 75 to 300 (Table 1). The difficulty indices are between 0.42-0.50 and are

in the desirable range. The discrimination indices are in the range of 0.24 to 0.35, which are in the good range. There was a statistically significant difference in the difficulty indices and the discrimination indices based on different subjects. The discrimination indices for microbiology and pharmacology were significantly higher than those for physiology, biochemistry, anatomy, and pathology. The mean iRT for each item was the shortest in parasitology (22.8 seconds), the longest in biochemistry (37.5 seconds), and there was a statistically significant difference based on different courses. The mean iRT for each item in physiology and biochemistry was longer than that for microbiology and pathology. Physiology, biochemistry, and anatomy had longer iRT than pharmacology. The iRT for physiology was longer than that for anatomy.

2. Correlation between response time and item analysis indices

There were statistically significant positive correlations between iRT, total test times, and total number of items, which is shown in Table 2. On the other hand, there were statistically significant negative correlations

Fig. 2. Scatter Plot of the Difficulty Index for Item Response Time

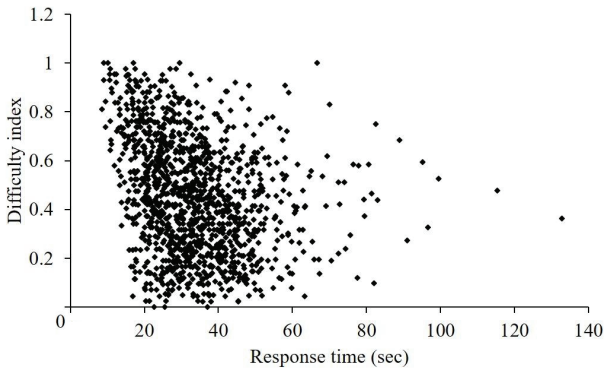


Table 1. Comparison of the Difficulty Index, Discrimination Index, and Item Response Time Based on the Courses (Year: 2013–2017)

Variable	Total no. of items	Difficulty index	Discrimination index	Response time (sec)
Microbiology[a1]	175 (35)	0.49±0.23	0.35±0.26	31.8±12.4
Pharmacology[a2]	175 (35)	0.48±0.21	0.35±0.24	30.6±17.3
Physiology[b1]	140 (35)	0.46±0.23	0.26±0.21	39.4±13.4
Biochemistry[b2]	175 (35)	0.45±0.24	0.24±0.24	37.5±16.9
Anatomy[b3]	300 (60)	0.42±0.23	0.25±0.24	35.2±10.7
Pathology[b4]	225 (45)	0.42±0.23	0.27±0.25	32.3±11.1
Parasitology[c]	75 (15)	0.50±0.23	0.27±0.17	22.8±7.4
F-value		3.303*	6.846**	18.629**
Multiple comparison			a1–2>b1–4	a1, b4<b1, b2 a2<b1, b2, b3 b1>b3 a, b1–4>c

Data are presented as number of items (mean) or mean±standard deviation, unless otherwise stated.
*Significant at the p<0.05 level. **Significant at the p<0.01 level.

Table 2. Intercorrelations for Item Response Time, Item Characteristic, and Classical Test Theory

Variable	1	2	3	4	5
1. Item response time	1				
2. Total test time	0.112**	1			
3. Item length	0.118**	0.999**	1		
4. Difficulty index	-0.263**	-0.102**	-0.101**	1	
5. Discrimination index	-0.145**	-0.067*	-0.066*	0.259**	1

*Correlation is significant at the 0.05 level. **Correlation is significant at the 0.01 level.

between iRT, item difficulty, and discrimination, which is shown in Figs. 2–4. This means that iRT increases as the difficulty of the items increases, and iRT decreases as the degree of discrimination increases. Difficulty indices and discrimination indices were also positively correlated.

3. Comparison of item response time according to difficulty and discrimination

iRT of the items with high difficulty was significantly longer in all of the courses except for pharmacology (Table 3). According to the degree of difficulty, iRT had a range of 13.2 seconds (parasitology) to 26.9 seconds (biochemistry) for items with low difficulty, and the items with high difficulty had an iRT from 26.8 seconds (parasitology) to 42.3 seconds (physiology). There was no statistically significant difference in iRT based on the level of item discrimination in the courses except for pharmacology. The mean iRT in pharmacology was 25.7 seconds for the items with high discrimination and 35.8 seconds for the items with low discrimination, which were statistically significant.

4. Relationship between item response time, difficulty, and discrimination indices

The regression analysis about difficulty and discrim-

Fig. 3. Scatter Plot of the Discrimination Index for Item Response Time

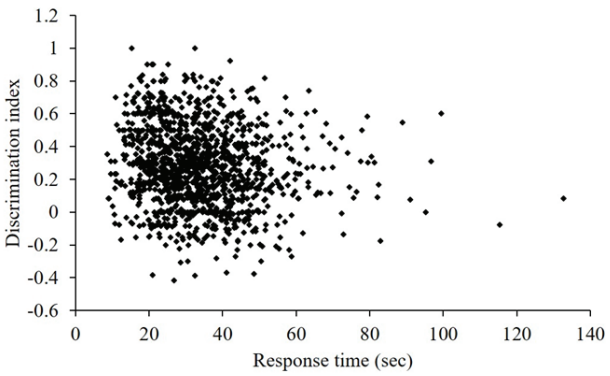


Fig. 4. Three-dimensional Scatter Plot of the Discrimination Index, Difficulty Index, and Item Response Time

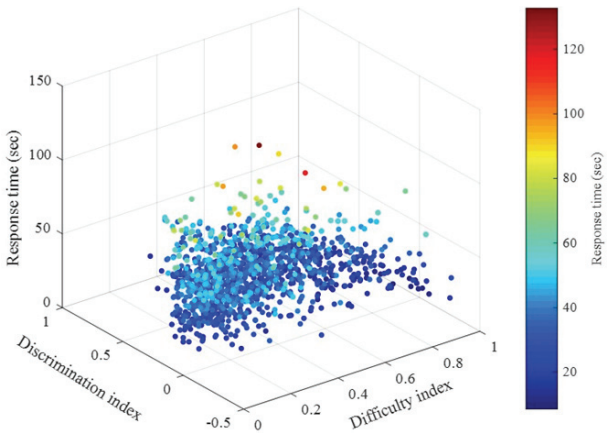


Table 3. Comparison of Item Response Times with Different Indices of Difficulty and Discrimination (Year: 2013–2017)

Variable	Anatomy	Pathology	Biochemistry	Microbiology	Pharmacology	Physiology	Parasitology
Discrimination index[a]							
Poor	131 (36.3)	94 (32.5)	78 (39.2)	58 (34.4)	51 (35.8)	53 (39.4)	28 (24.4)
Fair	47 (34.7)	30 (32.0)	23 (35.9)	15 (32.8)	24 (29.3)	31 (39.2)	17 (23.5)
Good	42 (33.9)	29 (34.5)	28 (37.3)	16 (31.8)	33 (33.4)	19 (41.3)	10 (18.0)
Very good	80 (34.2)	72 (31.2)	46 (35.4)	86 (29.9)	67 (25.7)	37 (38.6)	20 (22.2)
p-value	0.427	0.603	0.641	0.206	0.011	0.911	0.121
Difficulty index[b]							
Easy[1]	25 (25.2)	14 (26.0)	15 (26.9)	18 (22.2)	12 (27.9)	12 (23.7)	9 (13.2)
Desir[2]	167 (35.2)	136 (32.1)	106 (38.4)	116 (32.1)	125 (29.7)	89 (40.3)	52 (23.3)
Diffic[3]	108 (37.5)	75 (33.7)	54 (38.7)	41 (35.3)	38 (34.3)	39 (42.3)	14 (26.8)
p-value	0.000	0.056	0.037	0.001	0.306	0.000	0.000
Multiple comparison	1<2, 3	1<3	1<2, 3	1<2, 3		1<2, 3	1<2, 3

Data are presented as number of items (average). Year 2013 and 2015–2017 show p-values of 0.001 and 0.05, respectively.
[a]Poor: <0.2, fair: 0.2–0.29, good: 0.3–3.9, very good: >0.4. [b]Easy: >0.8, desirable: 0.3–0.8, difficult: <0.3.

Table 4. Regression Analysis for the Prediction of Response Time by Classical Test Theory Based on the Courses

Variable	All	Anatomy	Biochemistry	Physiology	Pathology	Microbiology	Pharmacology	Parasitology
Difficulty index	-0.143**	-0.134**	-0.129*	-0.220**	-0.099**	-0.181**	-0.105	-0.143**
Discrimination index	-4.658**	-3.041	-4.657	5.143	-1.142	-3.272	-13.131*	0.976

Data are presented as β. Overall model R^2=0.075, F=51.294.
*Significant at the p<0.05 level. **Significant at the p<0.01 level.

ination indices on iRT indicated that iRT increases as the difficulty level increases, and iRT decreases as the degree of difficulty increases, which is shown in Table 4. The overall model was significant, accounting for 7.5% of the variance in iRT. There was some variation according to the courses; iRT increased with increasing difficulty level in all of the courses except for pharmacology. In pharmacology, iRT showed no statistically significant relationship with the degree of difficulty; however, iRT decreased as the discrimination index increased, which was statistically significant.

Discussion

The results of analyzing the degree of item difficulty, discrimination, and iRT the seven courses of CBMSE for 5 years from 2013 to 2017 indicate that iRT increases with increasing degree of difficulty, and iRT tends to decrease with increasing degree of item discrimination. In other words, the students' effort is increased when they solve the difficult items but reduced when they are confronted with items with high degree of discrimination. The major findings of this study relative to iRT and item difficulty were consistent with earlier research. Response time increased with increasing item difficulty [6,7,9,10]. Previous studies indicated that examinees strategized by postponing choosing an answer if the item was too difficult for them to solve.

The CTT has limitations as the difficulty and discrimination of the items are different according to the group characteristics, and the examinees' ability is estimated differently according to the test characteristics [2]. In our study results, on the other hand, it was found that iRT tended to decrease with increasing item discrimination regardless of the course. Halkitsis et al. [9] found a positive correlation with discrimination and iRT in a study called "Licensing examination on microcomputers at Drake Authorized Testing Centers throughout the United States," which is different from our results. This is probably due to differences in the characteristics of the test. CBMSE is a relatively low-stakes evaluation test in Dankook University College of Medicine, which is different than a high-stakes test such as a qualification test.

Since higher discriminative item distinguishes students' ability to solve problems in those with higher grades and those with lower grades, this suggests that the students who are not prepared for an item may have engaged in a rapid-guessing behavior. Rapid-guessing is a response occurring so rapidly that examinees do not have time to fully consider the item [12]. In other words, it means that they gave up the item and marked an answer by guessing and therefore skipped it quickly to solve other items.

It is reported that in low-stake tests consisting of many items, a number of examinees engaged in solution behavior for most of the test (e.g., 40-50 items) and then abruptly switched to rapid-guessing behavior for the remainder of the items. In our study, the average number of items in the seven courses were between 15, 60, and 35-45, so we do not think that this is caused by a

rapid-guessing behavior due to the large number of items.

According to the test-taking model proposed by Wolf et al. [14], the amount of available energy appears to vary across examinees. In other words, the extent to which examinees devote their efforts to take an examination can vary according to the item difficulty and degree of discrimination. Wise and Kong [11] discovered that examinees who do not try to do well on a test item exhibit rapid-guessing behavior in low-stakes examination because examinees who are not motivated will respond quickly. Therefore, in a low-stakes test, it is necessary to keep the level of difficulty and degree of discrimination at an appropriate level so that as many students as possible can solve the problems with maximum effort. Adequate motivation of the students is as important as accurate assessment of their ability.

In our study, pharmacology items had a high degree of discrimination and short iRT compared to the other courses. In the final regression analysis, there was no correlation between item difficulty and iRT, while iRT had a strong negative correlation with degree of discrimination. How can we explain this difference? Although it is difficult to obtain a clear answer in this analysis, it can be speculated that the higher level of discrimination may cause the expression of stronger rapid-guessing behavior. In other words, if students perceive pharmacology to be a difficult course, such a difference can be caused by a lack of motivation in students, especially in those with low grades. For more accurate interpretation, further research such as a group-by-group analysis adjusted to different grades is required.

The limitations of this study are as follows. First, iRT used in this paper is defined as the time taken to make the final decision. It means that, if the student changes her final answer some time after the first marking, previous iRT for that item is deleted and replaced by a new one. In future studies, through the modification of the CBT software program, it may necessary to extract response time for each repeated problem-solving behavior for a given item. In that way, we can get multiple response times for a single item. There can be many response times we can define and extract from CBT to gain insight about the examinee's test-taking behavior. With the modification of the software program, we can calculate response time for items answered correctly and incorrectly. Response time for items with a 'wrong answer changed to a correct one' and a 'correct answer changed to a wrong one' and so on can be obtained too.

Second, previous work indicates that response time is negatively correlated with the total test time, iRT increases when there is a long item description or figure inclusion, and response time has a positive correlation with item difficulty [12]. In this study, we only considered difficulty, discrimination, and number of items without including variables that affect the response time of each test item, such as the length of the test items, pictures, and tables. Cognitive psychologists have focused on the within-person relationship between speed and accuracy. When a person chooses to perform a task more quickly, the person's accuracy tends to decline [15]. It may be necessary to analyze the relationship between response time and correct answer of the item in each examinee. Response time may also be used in the future to help identify unusual or cheating behaviors [16].

Third, the results of our study may have limited generalizability. The data came from just one college's CBT results. However, through the analysis of the seven courses over the past 5 years, we tried to explore the meaning of iRT in relation to CTT and presented some insights into the dynamics of students' problem-solving behavior. Despite the limitations of the study, our results suggest that appropriate difficulty and discrimination of

items can lead students to put their best effort into the test.

The conclusion of this study is as follows. There are negative correlations among item difficulty, degree of discrimination, and iRT; iRT decreased as item difficulty or the degree of discrimination increased. There was variation of this relationship depending on the course. The students' test effort as represented by iRT was properly maintained when the items had a 'desirable' difficulty level and a 'good' discrimination level. The CBMSE in Dankook University College of Medicine is a relatively low-stakes examination compared to other course examinations. To increase the students' motivation, an adequate degree of difficulty and discrimination power is required in such an examination. It may be inferred that with the combination of CTT and iRT, we can gain deeper insights about the quality of the examination and test behaviors of the students, which can provide us with more powerful tools to improve them.

ORCID:
Yoo-mi Chae: https://orcid.org/0000-0003-1071-6099;
Seok Gun Park: http://orcid.org/0000-0001-5824-6298;
Ilyong Park: https://orcid.org/0000-0003-1613-4209

Acknowledgements: None.

Funding: There was no financial support for this study.

Conflicts of interest: No potential conflict of interest relevant to this article was reported. The authors alone are responsible for the content and writing of the paper.

Author contributions: CBT data analysis and statistical results composition: YC; advising the research and the study contents: SP; CBT system, gathered data, and composed the paper: IP; and corrected the paper in English: YC, SP, IP.

References

1. Son CG. Curriculum & educational evaluation. Seoul, Korea; Teayong; 2012.
2. Lim HS, Lee YM, Ahn DS, Lee JY, Im H. Item analysis of clinical performance examination using item response theory and classical test theory. Korean J Med Educ. 2007;19(3):185-195.
3. University of Washington. Understanding item analyses. http://www.washington.edu/assessment/scanning-scoring/scoring/reports/item-analysis. Accessed July 3, 2018.
4. Huh S. Preparing the implementation of computerized adaptive testing for high-stakes examinations. J Educ Eval Health Prof. 2008;5:1.
5. Wise SL. An investigation of the differential effort received by items on a low-stakes computer-based test. Appl Meas Educ. 2006;19(2):95-114.
6. Lee YH. Introduction to an open source internet-based testing program for medical student examinations. J Educ Eval Health Prof. 2009;6:4.
7. Bergstrom BA, Gershon R, Lunz ME. Computerized adaptive testing exploring examinee response time using hierarchical linear modeling. Paper presented at: the Annual Meeting of the National Council on Measurement in Education; April 4-8, 1994; New Orlenas, USA.
8. Bovaird JAI. New applications in testing: using response time to increase the construct validity of a latent trait estimate. https://elibrary.ru/item.asp?id=5458647. Published 2004. Accessed July 3, 2018.
9. Halkitsis PN, Jones JP, Pradhan J. Estimating testing time: the effects of item characteristics on response latency. Paper presented at: the Annual Meeting of the American Educational Research Association; April 8-12, 1996; New York, USA.
10. Yang CL, O'Neill TR, Kramer GA. Examining item difficulty and response time on perceptual ability test

items. J Appl Meas. 2002;3(3):282-299.
11. Wise SL, Kong X. Response time effort: a new measure of examinee motivation in computer-based tests. Appl Meas Educ. 2005;18(2):163-183.
12. Wise SL. An investigation of the differential effort received by items on a low-stakes computer-based test. Appl Meas Educ. 2006;19(2):95-114.
13. Schoening A. Interpreting exam performance: what do those stats mean anyway? https://www.creighton.edu/sites/www12.creighton.edu/files/PtT-Exam%20Analysis.pdf. Accessed July 3, 2018.
14. Wolf LF, Smith JK, Birnbaum ME. Consequence of performance, test, motivation, and mentally taxing items. Appl Meas Educ. 1995;8(4):341-351.
15. Schnipke DL, Scrams DJ. Exploring issues of examinee behavior: insights gained from response-time analyses. In: Mills CN, Potenza MT, Fremer JJ, Ward WC, eds. Computer-Based Testing: Building the Foundation for Future Assessments. Mahwah, USA: Lawrence Erlbaum Associates; 2002:244.
16. Kingsbury GG, Zara AR, Houser RL. Procedures for using response latencies to identify unusual test performance in computerized adaptive tests. Paper presented at: the Annual Meeting of the National Council on Measurement in Education; April 13-15, 1993; Atlanta, USA.