https://doi.org/10.48009/1_iis_2017_128-136

Issues in Information Systems
Volume 18, Issue 1, pp. 128-136, 2017

A TOOL FOR PERFORMING ITEM ANALYSIS TO ENHANCE TEACHING AND LEARNING EXPERIENCES

Xue Bai, Virginia State University, xbai@vsu.edu
Ade Ola, Virginia State University, aola@vsu.edu

ABSTRACT

Assessment is one of the most essential educational tools. When properly developed and interpreted, assessments can help instructors better understand what their students are learning, identify specific areas which need clarification, and improve both teaching and learning experiences. Item analysis is a widely used tool for assessing learning outcomes. It provides the means for gathering evidence about students' knowledge on specific topics or knowledge areas. It can be used to identify students' strengths and weaknesses, to monitor student learning and progress, and to plan and conduct instruction. However, due to the cumbersomeness of calculating relevant statistics, some best practices in item analysis are too infrequently used in practice. This paper describes an item analysis tool that is integrated into the online programming assignment management and auto-grading system and presents results of an empirical study to show how it can be used to assess students' learning outcomes and enhance teaching and learning experiences.

Keywords: Item analysis, auto-grading, online programming assignment management

INTRODUCTION

Assessments are mainly used to measure learning outcomes, but the ultimate goal is to facilitate the learning process and improve learning outcomes. To get the most out of assessments, one needs to know what to assess and how to perform the assessment. To make assessment as effective as possible, assessment items (questions) must be well developed so that the results provide useful evidence about student learning and help instructors identify students' strengths and weaknesses. A poorly chosen or poorly developed assessment will fail to provide useful evidence about student learning; it could even provide misleading information. Therefore, one needs to ensure that all assessment items are appropriate for what's being assessed. A quality assessment item should not be too difficult or too easy, and it should be able to differentiate between students who do well on the overall test and those who do not. Item analysis is a relatively easy, yet valuable, procedure for assessing the quality of an assessment item. Item analysis uses statistics and expert judgment to evaluate a test based on the quality of individual items, item sets, and the entire set of test items, as well as the relationships among the test items (McCowan, 1995). Item analysis is a process which examines student responses to individual test items in order to assess the quality of each item and of the test as a whole. It evaluates how each item relates to the external criterion (the score of the test) and to the other items on the test (Thompson & Levitov, 1985, p.163). Information provided by item analysis can be used to improve item and test quality. Item analysis is especially valuable in improving items that will be used again in later tests, but it can also be used to eliminate ambiguous or misleading items in a single test administration. In addition, item analysis can help instructors improve their skills in test construction, and results of an analysis can be used to identify specific areas of course content that need greater emphasis or clarification.

STATISTICS IN ITEM ANALYSIS

Item analysis is a useful tool for assessing learning outcomes. It involves computing various statistics that can provide useful information for improving the quality and accuracy of multiple-choice, multiple selection, and true/false questions. Item analysis provides the means for gathering evidence about students' knowledge on specific topics or knowledge areas. It also provides data that can be used to identify students' strengths and weaknesses, to monitor student learning and progress, and to plan and conduct instruction. However, item analysis is not being used widely because of the extra effort, beyond assessment and grading of tests, that is required to calculate relevant statistics. To facilitate item analysis and promote its use in computer programming education, we have developed software that is integrated into an online programming assignment management and auto-grading system. The software computes and presents item analysis measures and other statistics automatically. Presented in the following sections are the item analysis statistics and measures generated by the software.

Item Difficulty

The proportion of participants that gets an item correct is called item difficulty, and it is usually denoted as *p* (Crocker & Algina, 1986). The larger the proportion, the easier the item. Item difficulty, *p*, is computed as:

$$p = \frac{R}{T} \qquad (1)$$

where p = item difficulty index
R = the number of students who answered the item correctly
T = the total number of students who answered the particular question

For a multiple choice question with a single correct option, the item difficulty is the percentage of students who answer the item correctly. For multiple selection questions, difficulty is defined as the percentage of students receiving the maximum score. For the computation of this statistic, items that are not attempted by students are treated as not administered and, thus, are not included in the computation. Item difficulty index ranges from 0% to 100%, and it is typically written as a proportion that ranges from 0.0 to 1.00. The higher the value, the easier the question. Item difficulty is relevant for determining whether students have learned the concept being tested. It also plays an important role in the ability of an item to discriminate between students who know the tested material and those who do not. The item will have low discrimination if it is so difficult that almost everyone gets it wrong or guesses, or so easy that almost everyone gets it right. Items that are too difficult or too easy do not help differentiate between students who scored high or low on the test as a whole. The discrimination factor will be further discussed in the next section.

Instructional Assessment Resources (IAR) acknowledged values of item difficulty index and their evaluation is shown in Table 1 (IAR 2011).

Table 1. Evaluation of Item Difficulty Index for Item Analysis

Item Difficulty Index(*p*)	Item Evaluation
Above 9.0	Very easy item
0.62	Ideal value
Below 0.20	Very difficult item

Source: Instructional Assessment Resources (IAR 2011)

Items with *p*-values above 0.90 are very easy and should be carefully reviewed based on the assessment purpose. For example, if the instructor is using easy “warm-up” questions or aiming for student mastery, then some items with p values above 0.9 may be warranted. In contrast, if the instructor is mainly interested in differences among students, these easy items may not be helpful. *P*-values below 0.2 indicate very difficult items and they should be reviewed for possible confusing languages, be removed from subsequent exams, or be identified as an area for re-instruction. If almost all of the students get an item wrong, there is either a problem with the item or students did not understand the concept. Items that are too easy or too difficult cannot discriminate adequately between student performance levels. In order to obtain maximum spread of student scores, it is best to use items with moderate difficulties. In general, moderate difficulty can be defined as the halfway point between perfect score and chance score. For example, for a four option item, moderate difficulty level is 0.62 because 100% is is perfect score and 25% is the chance score [25% + (100-25)%/2 = 0.625]. However, if the instructor is trying to determine the top percentage of students that learned certain concepts, highly difficult items may be necessary.

Item Discrimination

Item discrimination is used to measure the extent to which an item is a predictor of overall performance on a test. Item discrimination measures the discriminating power; it gives the relationship between the item response and the total test score for all participants. If an item is well designed, the students that obtain high scores on the test are more likely to get the question right. That is, students who answered the correct response scored well on the test, whereas students who did not answer correctly did not score well on the test. Item discrimination is computed as the correlation between a correct response to the item and the total score on the test. It is also referred to as the Point-Biserial correlation (PBS). The range of item discrimination value is from -1.0 to +1.0. The higher the value, the more discrimination the item provides. Items with discrimination values near or less than zero indicates that the students who did poorly overall on the test did better on that item than students who overall did well. In that case, the item may be confusing for your better scoring students in some way. Such item should be removed or redesigned. Item discrimination index (ID) is calculated as follows:

$$ID = \frac{(\overline{X}_C - \overline{X}_W)}{Std}\sqrt{p(1-p)} \quad (2)$$

where $\overline{X}_C$ = the mean total score for students who have responded correctly to the item
$\overline{X}_W$ = the mean total score for students who have responded incorrectly to the item
p = the item difficulty index for the item
Std = the standard deviation of the total exam scores

Ideally, an item should have a positive discrimination index of at least 0.20, which indicates that high scorers have a high probability of answering the item correctly and low scorers have a low probability of answering it correctly. Items can be classified with respect to item discrimination as shown in Table 2 (Ovwigho, 2013).

Table 2. Evaluation of Item Discrimination for Item Analysis

Item Discrimination Index(ID)	Item Evaluation
Above 0.4	Very good item
0.30 – 0.39	Reasonably good but subject to improvement
0.20 – 0.29	Marginal items and need improvement
Below 0.19	Poor items to be removed or redesigned

Source: Ovwigho (2013)

Difficulty index and discrimination index do have a correlation. Items that are too easy or too difficult will not be a good predictor of student performance. When calculating the discrimination index, the Point-Biserial correlation is maximized when p is near 0.50 (about half of the participants get it right). In practice, easy and difficult items should be removed from the assessment, or their number should be controlled. However, the mix of items should match the objectives of the assessment. For example, if the instructor is using easy "warm-up" questions or aiming for student mastery, then some items (10% of the items) with p values above 0.9 may be warranted. In contrast, if the instructor is mainly interested in differences among students, easy items may not be helpful.

Distractor Analysis

Distractors are the incorrect answers in a multiple choice question. Distractors impact greatly item difficulty and discrimination. In an effective instrument, each distractor is meant to uncover a particular understanding or misunderstanding in the students' thought process. According to Instructional Assessment Resources (IAR, 2011), student performance in an exam is very much influenced by the quality of the given distractors. Analyzing the distracters is useful in determining the relative usefulness of the decoys in each item. An Item should be revised if none of the students select certain multiple choice alternatives for that item (Matlock-Hetzel, 1997). Those alternatives are probably totally implausible and therefore provide little use as decoys; they do not provide any information to distinguish different levels of student performance. Millman and Greene (1993) suggested using discrimination index of each option to determine its usefulness as a distractor. The discrimination values for the distracters should be lower and, preferably, negative; whereas the discrimination value of the correct answer should be positive. This means that high scorers are less likely to select the distracters than low scorers. To demonstrate how each distractor is meant to uncover a particular understanding or misunderstanding in the students' thought process, we use one multiple choice item from an introductory java programming course to illustrate. The test item is as follows:

Which of the following statements about creating arrays and initializing their elements is false?

A. The new keyword should be used to create an array.

B. When an array is created with operator new, the number of elements must be placed in square brackets following the type of element being stored.

C. The elements of an array of integers have a value of null before they are initialized.

D. A for loop is commonly used to set the values of the elements of an array.

Table 3. Distractor Analysis

Option	A	B	C	D
No. of students	2	4	21[*]	0

[*]Denote the correct answer

Table 3 summarizes the number of selections for each distractor. The students who select distractor "A" understand how to specify array size in square bracket, but may not understand that an array is an object and must be created with the keyword: new. Students who selected distractor "B" may have a misunderstanding of how to specify an array size. Those who select the correct choice "C" understand how to create and initialize an array object. Nobody selected "D;" it is not a good distractor. Even though option "D" is not a good distractor, it did indicate that all students understood that a loop is commonly used to set the values of the elements of an array. On the other hand, too much distraction is also not a good decoy. Let us illustrate with Table 4, which is a summary of three items on a test.

Table 4. Sample Data

Question	A	B	C	D
#1	20*	0	2	3
#2	2	2	15*	6
#3	0	3*	19	3

* Denotes the correct answer

A good item will have this result: Students who received high score on the test will provide the correct response more frequently than those who got lower scores. If a distracter is not a plausible alternative and few or no students chose the alternative, or too many students, especially high scorers, chose the incorrect alternative instead of the correct response, then it does not help uncover any particular understanding or misunderstanding in the students' thought process. For example, in the table 4, no one selected choice B in question #1 or A in question #3; they are not good distractors. However, in question #3, the choice of C by most students may be an indication that the item is not a good distractor because it looks extremely plausible to most students. It may indicate that some contents need to be reinstructed.

Test Statistics

A capability is also provided to generate mean, median, and standard deviation for test score data. The basis statistics can provide useful information. The mean score gives a rough idea of how students performed as a whole. When a student's score is compared to the mean, we can say that a student's performance is above or below average of the whole class. The standard deviation gives an indication of how widely spread the scores are from the mean. A large standard deviation means that there is much variability in the test scores, and a number of students may be lagging behind. A small standard deviation means there is little variability among the scores; this may be an indication that majority of the students have similar understanding of the content being assessed.

Reliability

Reliability refers to consistency or stability of a measurement. Reliability is the extent to which test results are consistent, stable, and free of error variance. It is the extent to which a test provides the same ranking of students it is re-administered (McCowan, R., & McCowan, S., 1999). Reliability is considered the best single measure of test accuracy. Classic reliability is a test score that includes a true score and random error. A true score is defined as a theoretical, dependable measure of a student's obtained score uninfluenced by chance events or conditions (Spearman, 1904; SPSS, 1999). It is the average of identical tests administered repeatedly without limit. An instrument is considered reliable if it produces the same results every time it is used to evaluate identical measurement. Reliability is measured by Coefficient Alpha or KR-20. The following formula shows the reliability measured by Coefficient Alpha.

$$\alpha = \frac{k}{k-1}\left\{1 - \frac{\sum_{i=1}^{i=k} p_i(1-p_i)}{S^2}\right\}$$

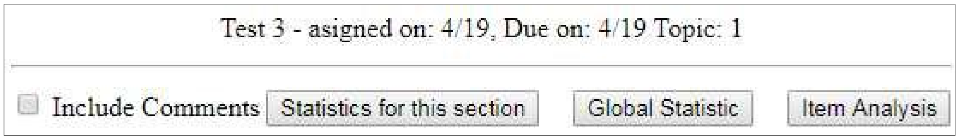
where k is the number of items on the exam; p_i, the item difficulty; and S^2 is the sample variance for the total score.

Reliability coefficients theoretically range in value from zero (no reliability) to 1.00 (perfect reliability). In practice, the acceptable reliability range is from .50 to .90. High reliability means that the questions on a test tend to "pull together." Students who answered a given question correctly were more likely to answer other questions correctly. If a parallel test were developed using similar items, the relative scores of students would show little change. Low reliability indicates that the questions tend to be unrelated to each other in terms of who answered them correctly. In those cases, the resulting test scores reflect peculiarities of the items or the testing situation more than students' knowledge of the subject matter. As with many statistics, it is dangerous to interpret the magnitude of a reliability coefficient out of context. High reliability should be demanded in situations in which a single test score is used to make major decisions, such as professional licensure examinations. Because classroom examinations are typically combined with other scores to determine grades, the standards for a single test need not be as stringent.

In the following section, we discuss the embedded item analysis tool that is integrated into the online programming assignment management and auto-grading system, and how it is used to assess students' learning outcomes and enhance teaching and learning experiences.

EMPIRICAL STUDY AND RESULT

Item analysis is a widely used tool for assessing learning outcomes. When properly developed and interpreted, item analysis can help instructors better understand what their students are learning, identify specific areas that need clarification, and improve both teaching and learning experiences. It provides the means for gathering evidence about students' knowledge on specific topics or knowledge areas. However, due to the cumbersomeness of calculating relevant statistics, some best practices in item analysis are too infrequently used in practice. We have developed a web-based item analysis system to facilitate the item analysis process. The item analysis system was developed using JavaServer Pages and Servlet, which is integrated into an online programming assignment management and auto-grading system. The system is a pure web-based application that is accessible from any type of browser. It provides an interface to allow instructors to perform item analysis with single click. Figure 1 shows the interface which is integrated with the online programming assignment management and auto-grading system.

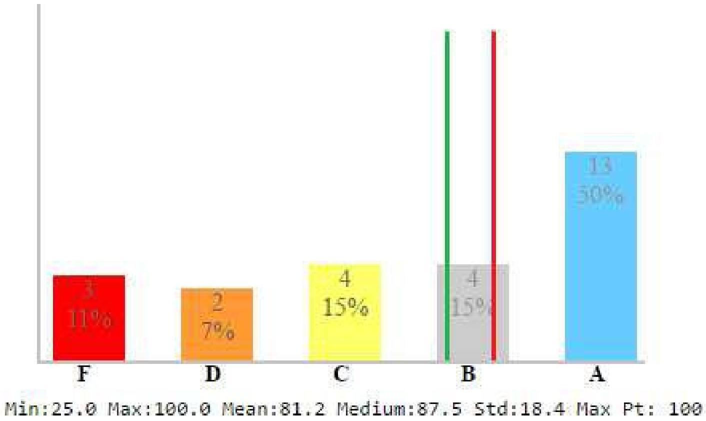


Test 3 - asigned on: 4/19, Due on: 4/19 Topic: 1

□ Include Comments

Figure 1. Integrated Tool for Performing Item Analysis

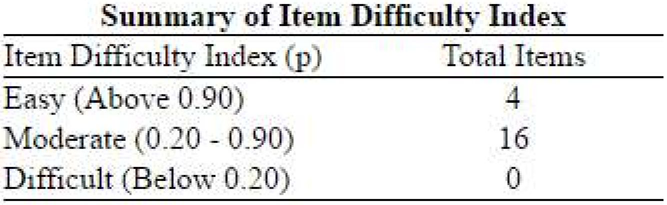
Several statistics can be used to describe the students' performance and test score distribution as a whole. These include mean score, median, min, max, range and standard deviation. While these statistics involve simple computations, instructors rarely use them to improve instruction and learning because of the extra effort required. The main contribution of the system development work reported in this paper is that it provides a one-click generation of basic statistics and item analysis measures. The software computes and presents statistics and item analysis automatically. The aim of the research presented is to automate item analysis computation and promote its use in computer programming education. The next four (4) figures show the statistics generated by the software for a test in Java programming course.



```
Min:25.0 Max:100.0 Mean:81.2 Medium:87.5 Std:18.4 Max Pt: 100
```

Figure 2. Statistic on a test

Figure 2 shows computed statistics for the test. The statistics include min, max, mean, medium, and standard deviation. The grade distribution is displayed as a histogram. The height of the bar for each grade category is proportional to the number of the students who received the grade in that category. The two color-coded vertical lines represent the locations of the mean and the medium: the green one (left one) is for mean and the red one is for medium. The data as shown in figure 2 indicates the grades are left skewed.

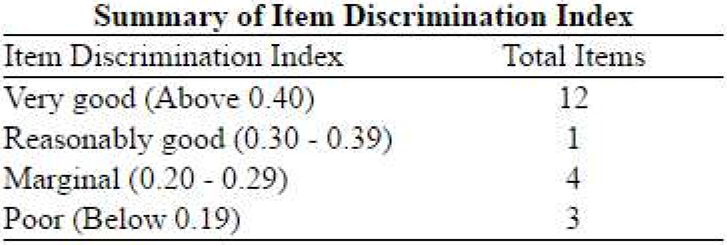


Summary of Item Difficulty Index

Item Difficulty Index (p)	Total Items
Easy (Above 0.90)	4
Moderate (0.20 - 0.90)	16
Difficult (Below 0.20)	0

Figure 3. Summary of Item Difficulty Index

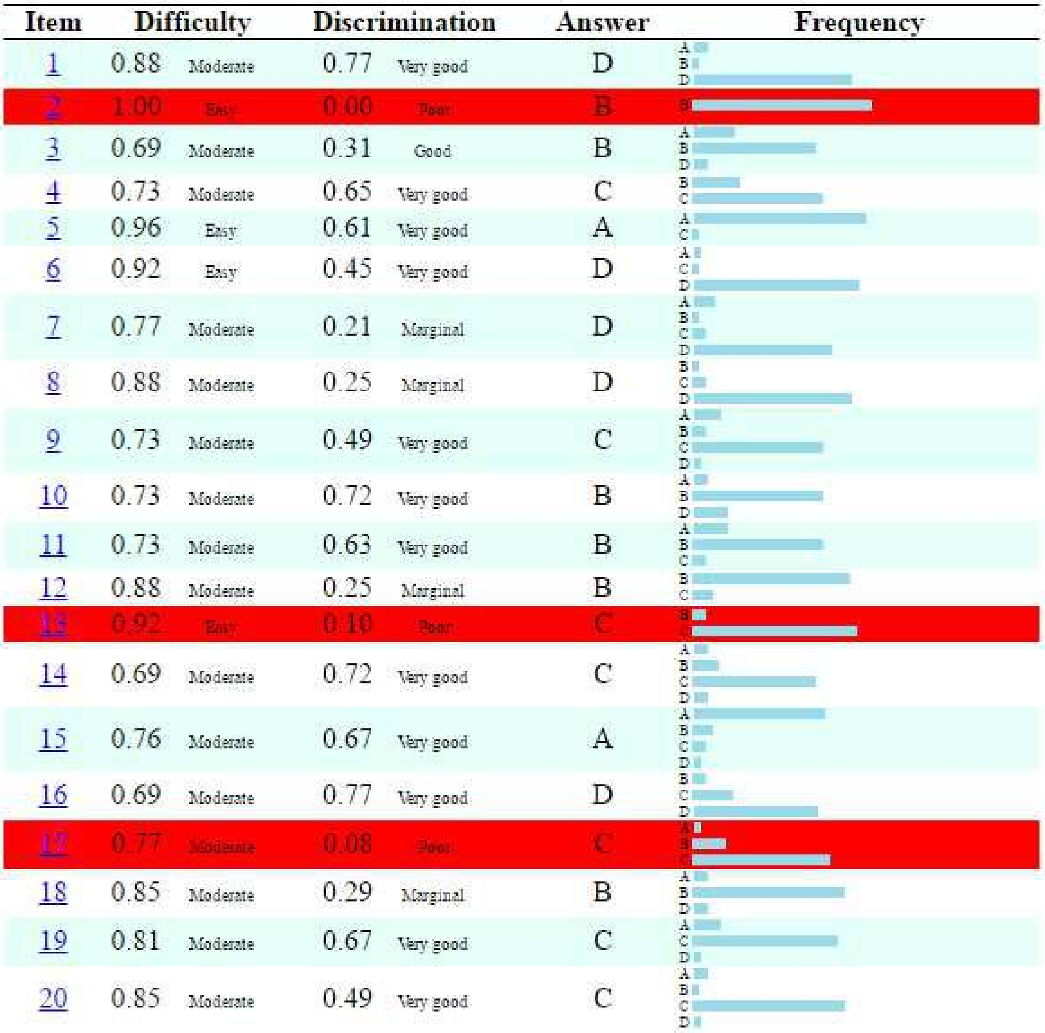
Figure 3 summarizes the difficulty index for the test. Based on the recommendations by the Instructional Assessment Resources (IAR), test items are classified into three categories in terms of difficulty index as shown in Figure 3. Figure 4 shows the item discrimination index for the test items. Based on the recommendations by Ovwigho (2013), the item discriminations are classified into four categories.



Summary of Item Discrimination Index

Item Discrimination Index	Total Items
Very good (Above 0.40)	12
Reasonably good (0.30 - 0.39)	1
Marginal (0.20 - 0.29)	4
Poor (Below 0.19)	3

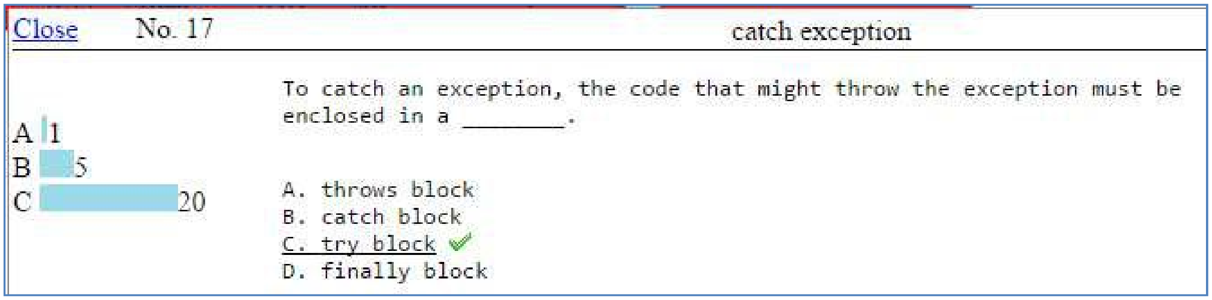
Figure 4. Summary of Item Discrimination Index



Item	Difficulty		Discrimination		Answer	Frequency
1	0.88	Moderate	0.77	Very good	D	A B D
2	1.00	Easy	0.00	Poor	B	B
3	0.69	Moderate	0.31	Good	B	A B D
4	0.73	Moderate	0.65	Very good	C	B C
5	0.96	Easy	0.61	Very good	A	A C
6	0.92	Easy	0.45	Very good	D	A C D
7	0.77	Moderate	0.21	Marginal	D	A B C D
8	0.88	Moderate	0.25	Marginal	D	B C D
9	0.73	Moderate	0.49	Very good	C	A B C D
10	0.73	Moderate	0.72	Very good	B	A B D
11	0.73	Moderate	0.63	Very good	B	A B C
12	0.88	Moderate	0.25	Marginal	B	B C
13	0.92	Easy	0.10	Poor	C	B C
14	0.69	Moderate	0.72	Very good	C	A B C D
15	0.76	Moderate	0.67	Very good	A	A B C D
16	0.69	Moderate	0.77	Very good	D	B C D
17	0.77	Moderate	0.08	Poor	C	A B C
18	0.85	Moderate	0.29	Marginal	B	A B D
19	0.81	Moderate	0.67	Very good	C	A C D
20	0.85	Moderate	0.49	Very good	C	A B C D

Figure 5. Statistics on Item Difficulty, Item Discrimination, and Frequency of Alternatives

Figure 5 shows the statistics on item difficulty index, item discrimination index, and frequency distribution of alternatives. Test items are listed according to their degrees of difficulty (easy, moderate, and hard) and discrimination (very good, good, marginal fair, poor). The distributions provide a quick overview of the test, and can be used to identify items in which students are not performing well and potential areas for improvement. In addition to the discrimination index for each alternative, the number and percentage of students who choose each alternative are reported (it is shown when the mouse cursor is moved over each frequency bar). The length of horizontal bars in the Frequency column represents the number of students who selected that alternative. Frequently chosen wrong alternatives may indicate common misconceptions among the students. The items with low discrimination index are highlighted with red. This information helps instructors identify specific areas for improvement. For example, the discrimination index of item 17 is 0.08, which may indicate some misconceptions among students. Clicking on an item will display detailed information on the item as shown in Figure 6.



No. 17 catch exception

```
To catch an exception, the code that might throw the exception must be
enclosed in a ________.

A. throws block
B. catch block
C. try block ✓
D. finally block
```

Figure 6. An Item with Poor Discrimination Index Value

From the detailed information, one can observe that some students misunderstood the difference between the code that might throw exception and the code to handle the exception if it did occur. This is an example of how to use the item analysis statistic to assess students' learning outcomes and enhance the teaching and learning experience.

SUMMARY

Item analysis can be a useful tool for instructors to gauge students' mastery of required skills and to monitor and evaluate teaching effectiveness and student learning. Through effective item analysis, instructors can identify specific areas of course content which need greater emphasis or clarification; this in turn will help improve the learning outcomes. However, item analysis is not being used widely because of the extra effort that is required to calculate relevant statistics. We have developed software that is integrated into an online programming assignment management and auto-grading system, to facilitate computation of item analysis statistics. The capability of the software is demonstrated with an empirical study that is presented in the paper. The aim of the research presented is to automate item analysis computation and promote its use in computer programming education.

REFERENCES

Crocker, L., & Algina, J. (1986). Introduction to Classical and Modern Test Theory. *New York: Holt, Rinehart and Winston*.

Instructional Assessment Resources, (2011). Item Analysis. Retrieved on March 20, 2017 from University of Texas at *Austin*, Instructional Assessment Resources, IAR Web site: *http://www.utexas.edu/academic/ctl/assessment/iar/students/report/itemanalysis.php*

Matlock-Hetzel, S. (1997). *Basic Concepts in Item and Test Analysis*. Retrieved on March 22, 2017 from http://ericae.net/ft/tamu/Espy.htm

McCowan, R. & McCowan, Sh. (1999). *Item Analysis for Criterion-referenced Tests*. Retrieved on March 8, 2017 from http://files.eric.ed.gov/fulltext/ED501716.pdf

Ovwigho, B .O. (2013). Empirical Demonstration of Techniques for Computing the Discrimination Power of a Dichotomous Item Response Test. *IOSR Journal of research and Method in Education, 3*(2), 12-17.

Sabri, S. (2013). Item Analysis of Student Comphrehensive test for Research in Teaching Beginner String Ensemble Using Model Based Teaching among Music Students in Public Universities, *International Journal of Education and Research, 1*(12), 1-13

Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology. 15,* 201-93.

SPSS. (1999). Item analysis. Spss.com Chicago: *Statistical Package for the social science*.

Thompson, B., & Levitov, J. E. (1985). Using microcomputers to score and evaluate test items. *Collegiate Microcomputer, 3,* 163-168.