

## ITEM ANALYSIS

### OBJECTIVE:

To increase understanding of item analysis and provide firsthand experience in calculation of difficulty ( $p$ ) and discrimination ( $d$ ) indexes.

### GENERAL INFORMATION:

After a test has been administered and scored, a *post hoc* analysis is often performed in order to evaluate the test's effectiveness. This procedure often involves an analysis of the individual items on the test. One primary goal of item analysis is to help improve the test by revising or discarding ineffective items. Another important function is to ascertain what test takers do and do not know.

In determining the usefulness of an item as a measure of individual differences in ability or personality assessment, the tester needs some external criterion measure of the characteristic. For example, if a test is being constructed to predict performance in a job or at school, then possible external criteria might be an index of job performance (such as supervisor's ratings) or an index of school achievement (such as grades assigned by teachers).

The validity of an item for predicting the particular external criterion measure may be determined by computing the correlation between scores on the item and scores on the external criterion measure. The most commonly used correlation coefficient in this context is the point-biserial coefficient.

With classroom achievement tests however, there is usually no external criterion against which items can be validated. Instead, another procedure is used that involves determining the percentage of test takers who passed each item and the correlation of each item with some criterion. In this case, though, the criterion consists of total scores on the test itself. Two statistics can help us to evaluate the usefulness of each test item.

The first is the **item-difficulty index ( $p$ )**. This index is determined by calculating the proportion of examinees that answer the item correctly. The second index is called the **item-discrimination index ( $d$ )**. For this calculation, we divide the test takers into three groups according to their scores on the test as a whole: an upper group consisting of the 27% who make the highest scores, a lower group consisting of the 27% who make the lowest scores, and a middle group consisting of the remaining 46%. (Note that some text books use 25%, not 27%.)

The formula for the item-difficulty index is

$$p = \frac{N_p}{N}$$

where:  $N_p$  indicates the number of test takers in the total group who pass the item, and  $N$  indicates the total number of test takers in the group

The formula for the item-discrimination index is

$$d = \frac{U_p - L_p}{U}$$

where:  $U_p$  and  $L_p$  indicate the numbers of test takers in the upper and lower groups who pass the item, and  $U$  is the total numbers of test takers in the upper group.

As an example, assume that 50 people take a test. For the difficulty index, 27 test-takers answers the item correctly. For the discrimination index, the upper and lower groups will be formed from the top 14 and bottom 14 test takers on total test score. If 12 of the test takers in the upper group and 7 of those in the lower group pass the item, then:

$$p = \frac{27}{50} = .54$$

$$d = \frac{12 - 7}{14} = 0.36$$

The item difficulty index ( $p$ ) has a range of 0.00 to 1.00. If no one answers the item correctly, the  $p$  value would be 0.00. An item that everyone answers correctly would have a  $p$  value of 1.00.

The optimal level for an acceptable  $p$  value depends on the number of options per item. A formula that can be used to compute the optimal level is:

$$\frac{1.0 + g}{2} \quad \text{where } g = \text{the chance level}$$

In our case, we will have four options; therefore,  $g = .25$ . Therefore, the optimal level for our tests will be .63. In general, as the number of options increases, the optimum  $p$  value decreases; we would expect questions with more options to also be more difficult to answer.

We can also compute the lower bound for item difficulty. We do not want too many questions to have a difficulty index below this bound, because it means that these items were too difficult for the group of examinees and therefore will not help us discriminate among the test takers.

$$\left[1 + 1.645\sqrt{(k-1)/n}\right] / k \quad \text{where } k = \text{number of multiple choice items}$$
$$n = \text{number of examinees}$$

For example, where  $k = 4$  and  $n = 90$ , the lower bound = .325.

$$\left[1 + 1.645\sqrt{3/90}\right] / 4 = .325$$

The item discrimination index ( $d$ ) is a measure of the effectiveness of an item in discriminating between high and low scorers on the whole test. The higher the value of  $d$ , the more effective the item is. When  $d$  is 1.00, all test takers in the upper group and no test takers in the lower group answered the item correctly. Conversely, if none of the upper group but all of the lower group answered an item correctly, the  $d$  value would be -1.00. Both of these circumstances are rare, and we will probably never see a value of -1.00. The range of values for the item discrimination index is -1.00 to 1.00. Generally speaking, an item is considered acceptable if its  $d$  index is 0.30 or higher.

#### YOUR HOMEWORK:

In this exercise, you are asked to calculate difficulty and discrimination indexes for results of two test items and to interpret the results.

Name \_\_\_\_\_

### ITEM ANALYSIS - WORKSHEET

#### SHOW ALL WORK

- 1) Compute the difficulty ( $p$ ) and discrimination ( $d$ ) indices of a test item administered to 84 people if 52 test-takers answered the item correctly; 20 in the upper group (upper 27% of total test score distribution) and 12 in the lower group (lower 27% of total test score distribution) got the item right? (Note:  $k = 4$ ) Is this a good item? Provide an interpretation for each of the calculated values.

$$p = \underline{\hspace{2cm}}$$

$$d = \underline{\hspace{2cm}}$$

---

---

---

- 2) Compute the difficulty( $p$ ) and discrimination ( $d$ ) index for an item administered to 263 people where 74 people answered the item correctly; 32 people in upper group and 23 people in the lower group passed the item ( $k = 4$ ). Is this a good item? Provide an interpretation for each of the calculated values.

$$p = \underline{\hspace{2cm}}$$

$$d = \underline{\hspace{2cm}}$$

---

---

---