Item Statistics Report

Admins
Item Administration. The number of times the item was administered (overall admins) or the number of times a particular item was marked (item admins).

$p$
Item difficulty. Determined by the correct response average and defined as the proportion of examinees who answered the item correctly. (Refer to Appendix A for a more detailed discussion of item difficulty.)

Avg
Average. Average overall test score for those examinees who answered the item correctly.

$p_{pb}$
Point biserial correlation. A statistical measure that summarizes how well a particular item predicted overall test success. (Refer to Appendix A for a more detailed discussion of point biserial correlation.)

Test Statistics Report

Median
The test score in the middle of the score distribution when all the scores have been ranked.

Mean
The arithmetic average of the test scores. The sum of the scores divided by the number of scores.

Standard Deviation
Statistical index that reflects the degree of dispersion in a group of scores. (Refer to Appendix A for a more detailed discussion of standard deviation.)

Test Reliability
Measure of how reliable a test is in differentiating student achievement. (Refer to Appendix A for a more detailed discussion of test reliability.)

Standard Error of Measurement
Statistical estimate of the difference between a examinee’s score on a test and his/her true score. Standard error of measurement is regarded as a constant property of the test and does not differ from one person to the next. (Refer to Appendix A for a more detailed discussion of standard error of measurement.)

Graph
Represents the distribution of test scores for all applicants. A normal distributed graph has a symmetrical bell curve and equal mean and median. A negatively skewed graph has a distribution with the majority of scores on the right side. A positively skewed graph has a distribution with the majority of scores on the left side. Skewed distributions may signify that the test included too few easy or too few hard items.

1 This document was compiled by Melissa B. Morris, Graduate Research Assistant at the IUPUI Testing Center.
Appendix A

Item difficulty ($p$)

Item difficulty is useful in evaluating whether test items are appropriate for the level of examinees taking the test. A lower score indicates that an item is more difficult than an item with a higher score. For example, an item with a difficulty of .2 is more difficult than an item with a difficulty of .7 because less examinees answered correctly to the former item. When $p$ is close to 0 or 1, the item is not providing any information about the differences between examinees’ abilities. A $p = 0$ means that every examinee answered the item incorrectly and a $p = 1$ means that every examinee answered the item correctly. Therefore, test developers should select items that have a range of difficulties that average about .5. Generally, item difficulties between .3 and .7 maximize the amount of information a particular test provides about differences among examinees. For true-false and multiple-choice items, the item difficulty should be adjusted to account for the effects of guessing. The optimal item difficulty for true-false items is halfway between .50 and 1.00 and for a four-option multiple-choice item, the optimal item difficulty is about .63. Generally, the optimal item difficulty can be calculated from the formula $(1.0 + g)/2$ where $g$ is the chance success level. For example, a four-option multiple-choice item’s optimal item difficulty would be $(1.0 + .25)/2$ or .63.

Point biserial Correlation ($r_{pb}$)

Point biserial correlation indicates the degree to which responses on one item are related to responses on other items. It measures the discrimination power of a particular item, i.e., how closely related success on an item is to success on the overall test. Assuming that total test scores accurately discriminate among examinees, a high positive $r_{pb}$’s for the correct responses present the most discriminating items. Therefore, examinees that answered the item correctly scored well on the tests, whereas examinees that answered the item incorrectly scored poorly on the test. On the other hand, a high negative correlation is good for the $r_{pb}$ of a distractor item, i.e., examinees that answer with an incorrect alternative did not score well on the total test. Point biserial correlation is an important indicator to consult in determining which items are discriminating and should be retained or revised. A good discriminator has a $r_{pb}$ of .40 or above. The formula used to compute $r_{pb}$ is:

$$r_{pb} = \frac{\bar{Y}_1 - \bar{Y}}{S} \sqrt{\frac{n_1 n}{(n-n_1)(n-1)}}$$

where:
- $n$ = total number of students
- $n_1$ = number who answered the question correctly
- $\bar{Y}$ = mean score for all students
- $\bar{Y}_1$ = mean score for those who answered the question correctly
- $S$ = standard deviation

Standard Deviation ($S$)

Standard deviation ($S$) summarizes the distribution of scores on a particular test. If the standard deviation is small, the scores are concentrated around a central value. When the scores are more spread out, the standard deviation becomes larger. The larger the standard deviation, the better the test is at discriminating among examinees’ performance levels. The formula used to calculate the standard deviation is:
\[
S = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}
\]

where:
\( n \) = number of students  
\( X_i \) = student’s score  
\( \bar{X} \) = test mean

**Standard Error of Measurement (SEM)**

Standard error of measurement (SEM) estimate of the probable difference between an examinee’s score on a test and his/her true performance. SEM is an index of measurement error and is present in every test. When the SEM is larger, the test is a less precise measure of examinee ability and less reliable. Two-thirds of the time examinee’s obtained scores fall within one SEM of their true score. The formula used to calculate the standard error of measurement is:

\[
SEM = s\sqrt{1 - r_{xx}}
\]

where:
\( r_{xx} \) = test reliability estimate

**Test Reliability**

Coefficient alpha (\( \alpha \)) is an index of the internal consistency of the test items (i.e., the tendency of the test items to correlate positively with one another). This reliability index tells test developers whether the test items tend to show a consistent interrelatedness. The coefficient alpha can vary between from 0.0 to 1.0. A reliability of .70 or higher is desirable for classroom tests. The formula used to calculate test reliability (coefficient alpha) is:

\[
\alpha = \frac{b}{b-1} \left[ 1 - \frac{\sum_{i=1}^{b} S_i^2}{\sum_{i=1}^{b} S_i^2} \right]
\]

where:
\( \alpha \) = coefficient alpha (estimated internal consistency reliability)  
\( b \) = total number of questions on the test

\[
S_i^2 = \frac{1}{n-1} \sum_{j=1}^{n} (X_{ij} - \bar{X}_i)^2  \quad \quad  S_i^2 = \frac{1}{n-1} \sum_{j=1}^{n} (t_j - \bar{t})^2
\]

or

\[
S_i^2 = SD^2
\]

(Note: \( S_i^2 \) is the “item score variance,” and \( S_i^2 \) is the “total test score variance.”)

where:
\( n \) = number of students  
\( t_j \) = test scores for each student  
\( j \) = a particular student  
\( t \) = test mean  
\( X \) = student’s score on a question  
\( \bar{X} \) = student average
References:


Contributor Notes

Correspondence concerning this document should be addressed to Dr. Howard R. Mzumara, Director, IUPUI Testing Center, 620 Union Drive, Room 003S, Indianapolis, IN 46202-5167. E-mail may be sent to hmzumara@iupui.edu.