Item analysis

An item analysis involves many statistics that can provide useful information for improving the quality and accuracy of multiple-choice or true/false items (questions). Some of these statistics are:

**Item difficulty:** the percentage of students that correctly answered the item.

- Also referred to as the p-value.
- The range is from 0% to 100%, or more typically written as a proportion of 0.0 to 1.00. The higher the value, the easier the item.
- Calculation: Divide the number of students who got an item correct by the total number of students who answered it.
- Ideal value: Slightly higher than midway between chance (1.00 divided by the number of choices) and a perfect score (1.00) for the item. For example, on a four-alternative, multiple-choice item, the random guessing level is 1.00/4 = 0.25; therefore, the optimal difficulty level is .25 + (1.00 - .25) / 2 = 0.62. On a true-false question, the guessing level is (1.00/2 = .50) and, therefore, the optimal difficulty level is .50+(1.00-.50)/2 = .75
- P-values above 0.90 are very easy items and should be carefully reviewed based on the instructor's purpose. For example, if the instructor is using easy “warm-up” questions or aiming for student mastery, then some items with p values above .90 may be warranted. In contrast, if an instructor is mainly interested in differences among students, these items may not be worth testing.
- P-values below 0.20 are very difficult items and should be reviewed for possible confusing language, removed from subsequent exams, and/or identified as an area for re-instruction. If almost all of the students get the item wrong, there is either a problem with the item or students were not able to learn the concept. However, if an instructor is trying to determine the top percentage of students that learned a certain concept, this highly difficult item may be necessary.

**Item discrimination:** the relationship between how well students did on the item and their total exam score.

- Also referred to as the Point-Biserial correlation (PBS) The range is from –1.00 to 1.00.
- The higher the value, the more discriminating the item. A highly discriminating item indicates that the students who had high exams scores got the item correct whereas students who had low exam scores got the item incorrect.
- Items with discrimination values near or less than zero should be removed from the exam. This indicates that students who overall did poorly on the exam did better on that item than students who overall did well. The item may be confusing for your better scoring students in some way.
- Acceptable range: 0.20 or higher
- Ideal value: The closer to 1.00 the better

\[
\text{Calculation: } \frac{\overline{X}_C - \overline{X}_T}{\text{S.D.Total}} \sqrt{pq}
\]

where

\[
\overline{X}_C = \text{ the mean total score for persons who have responded correctly to the item}
\]

\[
\overline{X}_T = \text{ the mean total score for all persons}
\]

\[
p = \text{ the difficulty value for the item}
\]

\[
q = (1 - p)
\]

\[
\text{S. D. Total} = \text{ the standard deviation of total exam scores}
\]
Reliability coefficient: a measure of the amount of measurement error associated with a exam score.

- The range is from 0.0 to 1.0.
- The higher the value, the more reliable the overall exam score.
- Typically, the internal consistency reliability is measured. This indicates how well the items are correlated with one another.
- High reliability indicates that the items are all measuring the same thing, or general construct (e.g. knowledge of how to calculate integrals for a Calculus course).
- With multiple-choice items that are scored correct/incorrect, the Kuder-Richardson formula 20 (KR-20) is often used to calculate the internal consistency reliability.

\[
\frac{K}{K-1} \left(1 - \frac{\sum pq}{\sigma^2_x}\right)
\]

K = number of items
p = proportion of persons who responded correctly to an item (i.e., difficulty value)
q = proportion of persons who responded incorrectly to an item (i.e., 1 – p)
\(\sigma^2_x\) = total score variance

- Three ways to improve the reliability of the exam:
  1) increase the number of items in the exam,
  2) use items that have high discrimination values in the exam,
  3) or perform an item-total statistic analysis

- Acceptable range: 0.60 or higher
- Ideal value: 1.00

Distractor evaluation: Another useful item review technique to use.

The distractor should be considered an important part of the item. Nearly 50 years of research shows that there is a relationship between the distractors students choose and total exam score. The quality of the distractors influences student performance on an exam item. Although the correct answer must be truly correct, it is just as important that the distractors be incorrect. Distractors should appeal to low scorers who have not mastered the material, whereas high scorers should infrequently select the distractors. Reviewing the options can reveal potential errors of judgment and inadequate performance of distractors. These poor distractors can be revised, replaced, or removed.

One way to study responses to distractors is with a frequency table. This table tells you the number and/or percent of students that selected a given distractor. Distractors that are selected by a few or no students should be removed or replaced. These kinds of distractors are likely to be so implausible to students that hardly anyone selects them.

- Definition: The incorrect alternatives in a multiple-choice item.
- Reported as: The frequency (count), or number of students, that selected each incorrect alternative
- Acceptable Range: Each distractor should be selected by at least a few students.
- Ideal Value: Distractors should be equally popular
- Interpretation:
  - Distractors that are selected by a few or no students should be removed or replaced
  - One distractor that is selected by as many or more students than the correct answer may indicate a confusing item and/or options
- The number of people choosing a distractor can be lower or higher than the expected because:
  - Partial knowledge
  - Poorly constructed item
  - Distractor is outside of the area being tested
Additional Information


Field, A. *Research Methods II: Reliability Analysis*. The University of Sussex.


Yu, A. *Using SAS for Item Analysis and Test Construction*. Arizona State University,

Modified from the original document housed at the Faculty Innovation Center, University of Texas at Austin:
facultyinnovate.utexas.edu