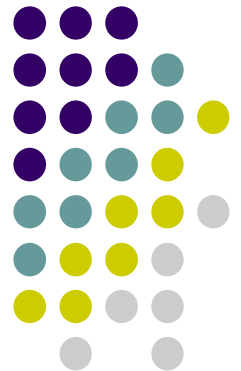


University Testing Services



Exam Scores: How to Interpret your Statistical Analysis Reports



Arizona State University

University Testing Services

Payne Hall, Room 301

480-965-7146

Interpreting the Item Analysis Score Report Statistical Information

This guide will provide information that will help you interpret the statistical information relating to the Item Analysis Report generated by University Testing Services when a multiple-choice exam is machine scored. It is important to note that three types of information are critical in understanding the validity of a multiple-choice exam.

These three are:

- 1) Whether the items were too difficult or too easy.
- 2) Whether the items discriminated between those students who really knew the material and those that did not.
- 3) Whether the incorrect responses in fact “distract” from the correct response or have no value whatsoever.

Several terms will be referenced in this guide. They include:

- 1) The Difficulty Factor
- 2) The Discrimination Index
- 3) The Kuder Richardson Reliability Coefficients
- 4) Differential Question Weighting
- 5) The Mean, Median, and Standard Deviation

The Difficulty Factor

The Difficulty Factor of a question is the proportion of respondents selecting the right answer to that question. It is a measure of how difficult the question was to answer. The following formula is used to calculate this factor.

$$D = c / n$$

D - Difficulty factor

c - Number of correct answers

n - Number of respondents

The higher the difficulty factor-the easier the question is. A value of 1.000 means that all of the students answered this correct response and this question may be too easy. The range of item difficulties on a good test depends on what you wish to know. If the purpose of a test is to determine if the students have mastered a topic area, high difficulty values should be expected. If the purpose of a test is to discriminate between different levels of achievement, items with difficulty values between 0.3 and 0.7 are most effective. **The optimal level should be 0.5.**

The Discrimination Index

The Discrimination Index measures the extent to which item responses can discriminate between individuals who have a high score on the test and those that get a low score. This is calculated for each response. The value is calculated with the following formula.

$$DI = (a - b) / n$$

DI - Discrimination Index

a - Response frequency of the upper quartile (75th percentile and above)

b - Response frequency of the lower quartile (25th percentile and below)

n - Number of respondents in the upper quartile (75th percentile and above)

A negative value means that students receiving a low score tended to select the option more than higher-scoring students. Conversely, a positive value for this index means that higher-scoring students tended to select the response more often. Ideally, the correct response should have a positive. A value of 0.0 indicates that there was no difference between the two groups. An item's difficulty can affect the discrimination index. Items which are very easy or very difficult will not discriminate very well between high- and low- scoring groups. On these items, nearly everyone will have gotten the item right or wrong regardless of how they performed on the other items on the test. Items which discriminate well are those which have difficulties between 0.3 and 0.7. **The value should be in a positive direction for a correct response and a negative direction for an incorrect response. If a response has a zero value the response should be eliminated as a choice.**

Kuder Richardson 20

This statistic measures test reliability of inter-item consistency. A higher value indicates a strong relationship between items on the test.

The KR 20 is calculated as follows:

$$KR = \frac{N}{N-1} * \frac{V - \sum (p_i q_i)}{V}$$

KR - Kuder Richardson 20

N - Number of items in the test

V - Variance of the raw scores or standard deviation squared

p_i - Proportion of correct answers of question i, or (number of correct answers /total number of responses)

q_i - Proportion of incorrect answers of question i, or (1 - p)

A lower value indicates a weaker relationship between test items. Values range from 0 to 1. **The better tests are within the .80 to .85 range.**

The Kuder Richardson 20 will always be greater or equal to the Kuder Richardson 21.
It is considered more accurate than the Kuder Richardson 21.

Kuder Richardson 21

This statistic approximates inter-item consistency. A high value indicates a strong relationship between items and a lower value a weaker relationship. The mean or average score of the exam is part of the formula. The values also range from 0 to 1.

The formula is calculated as follows:

$$\mathbf{KR} = \frac{N}{N-1} * \frac{1-[M(N-M)]}{N*V}$$

KR - Kuder Richardson 21

N - Number of items in the test

M - Arithmetic mean of the test scores

V - Variance of the raw scores or the standard deviation squared

Better than average tests should have a slightly lower value than the KR 20.

Differential Question Weighting

An example of a differential weighting system would be to give two points credit to some questions because they are more difficult or more important and one point credit to others. Reasons not to use this system include:

1. That there is no increase in reliability when using differential weighting
2. Differential weighting is most effective only in tests that are short (less than ten items).

Other Terms to Understand When Reviewing Score Reports-

Mean Score- Average score

Median Score- The score corresponding to the 50th percentile. Exactly half the scores are lower and higher.

Standard Deviation- The range above or below the average score where the majority of the scores lie. The more the scores are spread out, the higher will be the standard deviation.

References

ExamSYSTEM II User's Guide. (2000) Minneapolis, MN:National Computer Systems, Inc.

Interpreting Your Statistical Analysis. (1997) Provo, UT:BYU Testing Services

Lewis, Ralph F. & Ortiz, Kenneth K. (1988) *All of the above...A Guide to Classroom Testing and Evaluation* (2nd edition). Costa Mesa, CA: Economics Research, Inc.