



Exam Scores: How to Interpret your Statistical Analysis Reports



Arizona State University

University Testing and
Scanning Services
1130 E University, Suite 204
Tempe, AZ 85287-5204
480.965.7146
480.965.6859 (Fax)
<http://uoeee.asu.edu>

Interpreting the Exam Score Reports' Statistical Information

This guide will provide information that will help you interpret the statistical information relating to the reports 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
6. The Raw Score
7. The Stanine
8. A Subtest
9. The T-Score
10. The Z-Score

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. A value of 1.0000 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 be discriminated 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 discriminated 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:

$$KR = N$$