Understanding the TSAR output reports

Basic Item Response Data

This report shows the distribution of answers for multiple choice (or T/F) tests. For each item, the point value for the item is given, the correct answer, the number of students choosing each answer, the number of students who got the item right or wrong. The **mean** (average score) for each item is given, which is dependent on the point value for that item. The **variance** and **standard deviation** (square root of the variance) indicate how much variability there is in the responses for that item. If everyone chose the same answer, the variance and standard deviation would be 0.

What to look for: Look down the **Mean** column to see which items were most or least likely to be answered correctly. Look down the **Standard Deviation** column to see which items produced the most variability in response. The ones with higher standard deviations had the most students choosing different responses. Low standard deviations mean most students chose the same response.

Breakdown by Total Item Score

This report groups students by their total test score - lowest 27%, highest 27%, and the middle group (46%) - and shows which responses each group selected for each item. There is a summary column for the number in the low and high groups who got the item right or wrong.

The **Diff Index** (item difficulty index) is really the 'easiness' index. It shows the percent of all students who answered the item correctly. (It is the same as the **mean** on the above report, but with a different decimal place.) The difficulty index ranges from 0 to 100; the higher the index, the more students answered that item correctly. *What is the optimum value for the Diff Index?* That depends on the purpose of the test. Items that are intended to be challenging and identify the best students will have a low index. On the other hand, items that everyone is expected to know should have a high index.

The **Disc Index** (item discrimination index) is a measure of the effectiveness of an item in discriminating between high and low scores on a test. The notion is that high-achieving students will tend to choose the right answer, and low-achieving students will tend to choose the wrong answer. Item discrimination attempts to measure differences that truly exist among test takers. The Disc value is essentially the proportion of test takers in the highest group who chose the right answer minus the proportion in the lowest group who chose the right answer. When the Disc Index = 1, it means that all test takers in the high group answered the item correctly, and no test takers in the low group did. The closer the Disc Index is to 1, the better the item discrimination. When Disc Index = 0, the same proportion of test takers in the high and low groups answered correctly. If the index is negative, it means that proportionately more students in the low-achieving group answered that item correctly than in the high-achieving group. Items with negative values should be reviewed; the item was probably misleading or ambiguous.

What to look for: Item discrimination is greatly affected by item difficulty. Item discrimination is best when item difficulty is around 50. As a rule, Disc values of .20 or higher are desirable. Use this table as a guide to item quality:

Diff Index	*Disc Index*	*Decision*
35 to 85	.20 or higher	Ideal type of item
35 to 85	Below .20 or negative	Item does not discriminate and should be removed or revised.
Over 85	Irrelevant	Item is very easy. Retain only if the item measures essential material that everyone must know.
Under 35	.20 or higher	Although this item is hard, it does discriminate and should be retained, but used sparingly.
Under 35	Below .20	This item performs so poorly that it should be dropped or revised.

Analysis of distractors: The incorrect choices for an item are called distractors. Count the number of times that each distractor is selected by the upper group and the lower group. Ideally, all distractors should be equally plausible to examinees who don't know the right answer. Eliminate distractors that are never chosen; they are not working. Look into distractors most often selected by the high-achievers; students may be reading something into them that wasn't intended.

Statistics for Total Item Score

This report summarizes scores on the test for all students: **Mean** (average – depends on the point values of the items); **Median** (if you order the scores from lowest to highest, it is the halfway point); and measures of overall variability (Variance, Standard Deviation and Standard Error of the Mean).

The **Kuder-Richardson 20 and 21** are measures of internal consistency of the test – to what extent do the items provide consistent information about the students' level of knowledge as assessed by the test? Assuming that all the items on a test relate to a single content domain, we would expect students with a very high level of knowledge in the domain to answer most items correctly, and students with a low level to answer incorrectly. K-R 20 is more accurate than K-R 21 (which assumes that all items are of equal difficulty.) Values range from 0 to 1, with values closer to 1 reflecting greater consistency. Values over .70 are generally considered acceptable; however, if a test covers more than one content area, values can be lower. Also, tests that are criterion- (rather than norm) referenced are not candidates for reliability analysis. NOTE: If any test item has a point value greater than 1, ignore the Kuder-Richardson coefficients. The TSAR formula is incorrect for these items.

Prepared by: Sharron Ronco, Assessment Director
Sharron.Ronco@Marquette.edu
February, 2013