ITEM ANALYSIS

Multiple-Choice Questions: Difficulty and Discrimination Indices

Good multiple-choice questions (items) follow the Goldilocks rule: they should not be too difficult or too easy, but just right. Furthermore, they should appropriately discriminate between students who really know the answer and students who do not. In other words, they reflect a student's true score. Unfortunately, the true score is theoretical; we cannot really know what it is supposed to be. However, we can create good test items that mitigate error so that a student's observed score is as close to their true score as possible. So how do we know if our test questions are good?

An item analysis allows one to determine whether a multiple-choice question discriminates between students who know the material from those who do not and consists of calculating two indices for each question: a difficulty index and a discrimination index (Salkind, 2017).

Difficulty Index (D) – describes the total number of students who got an item correct.

$D = (S_H + S_L)/T$

Where

D = difficulty index
S_H = number of students in the high group (see below) who answered the question correctly
S_L = number of students in the low group (see below) who answered the question correctly
T = the total number of responses for the item

Interpreting the difficulty index requires students to be divided into high and low groups.

High and low groups – defined by the top and bottom 50% of total exam scores. To calculate the high and low groups:

1. For each student, tally the total number of multiple-choice questions that they got correct.

2. Rank and sort the score totals (still linked to each student) from highest to lowest.

3. Identify the top 50%* of scores as the high group, and the bottom 50% of scores as the low group. If you have an odd number of students who took the exam, use your judgement of whether the median score best fits into the high or low group. (*Much of the literature will say to use the top 27%, but using the top 50% is OK for smaller class sizes.)

Ideally, if 50% of the students are the high group, and 50% are in the low group, then the difficulty index should also be 50%, meaning half got it right and half got it wrong. Ideally as well, the top 50% scorers are the ones who got it right.

The more the difficulty index differs from 50%, the more poorly the item discriminates. Indices higher than 50% suggest that the item may be too easy. Indices lower than 50% suggest that the item may be too hard.

Discrimination Index (d) – describes how effectively an item discriminates between the high and low groups of students.

$d = (S_H - S_L)/(0.5*T)$

Where

d = discrimination index
S_H = number of students in the high group who answered the item correctly
S_L = number of students in the low group who answered the item correctly
T = the total number of responses for the item

Similar to a correlation coefficient, d ranges from 0.0 to +/- 1.0, where 0.0 indicates no discrimination and 1.0 indicates perfect discrimination. When d is positive and closer to 1.0, more students in the high group got the item correct, so the item is discriminating like it is supposed to. When d is negative and closer to -1.0, more and more students in the low group got the item correct, so the item is not doing what it is supposed to do.

In all likelihood we won't achieve perfect difficulty and discrimination indices, so what is a good multiple-choice question? Here's some perspective from Salkind (2017), shown in Figure 1. According to the difficulty index (horizontal), 25% of students got question 1 (Q1) correct. It may be a little difficult, but it positively discriminates between students in the high and low groups, which is a good thing. Seventy-five percent of students got Q2 correct. It positively discriminates buy may be a little too easy. Similar to Q1, 25% of students got Q3 correct, but it does not discriminate between the high and low groups, which is not good. Q4 is one we would definitely want to think about throwing out and rewriting. It has a 75% difficulty index, so it is probably too easy, and it negatively discriminates, meaning more students in the low group than the high group got it correct.

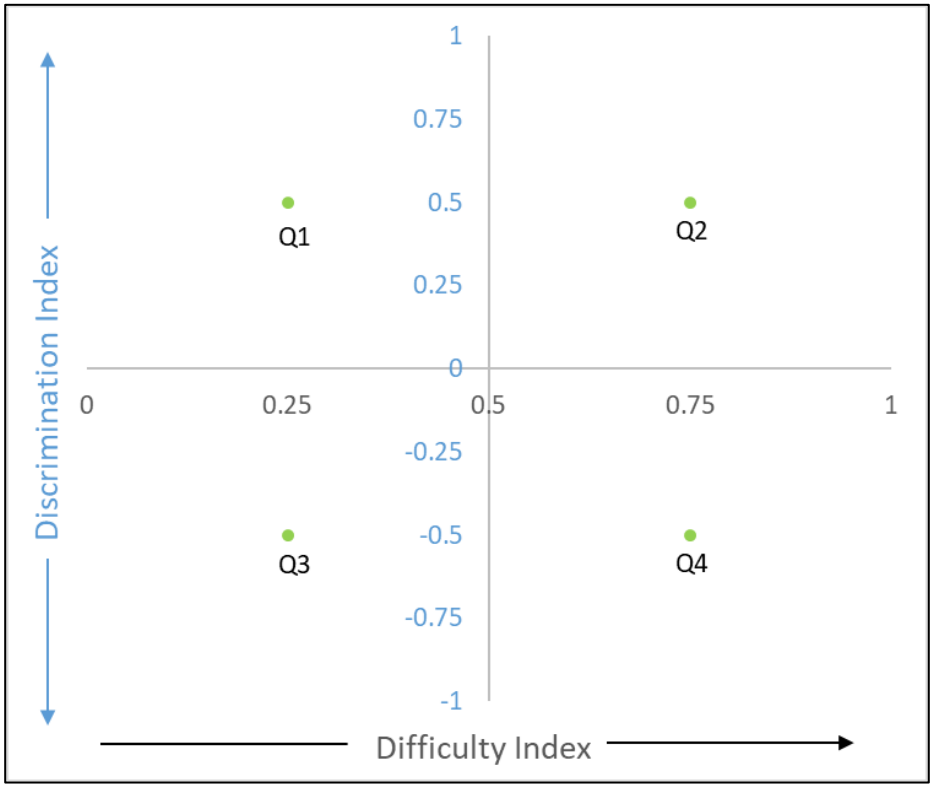

Figure 1. Difficulty and Discrimination Indices (adapted from Salkind, 2017).

Reference:

Salkind, N. J. (2017). Tests & measurement for people who (think they) hate tests & measurement. 3rd ed. Thousand Oaks, Calif: SAGE Publications. p. 161-172