Shiken: JALT Testing & Evaluation SIG Newsletter, 3 (2) Jan. 2000 (p. 18 - 20)

Statistics Corner

Questions and answers about language testing statistics:

Questions and answers about language testing statistics: item statistics for weighted items

James Dean Brown (University of Hawai'i at Manoa)

QUESTION: I have read your testing book (Brown, 1995, or see 1999 for a Japanese version of that same book), and I understand how to do item analysis for norm-referenced tests using item facility and item discrimination statistics and for criterion-referenced tests using the difference index and B-index. But all of the examples in your book are for tests that are scored right/wrong with the students' answers coded "1" for right and "0" for wrong. My question is: What can you do when you want to do item analysis for items that have weighted scores instead of right/wrong scorings? (This question was raised by Dr. Kimi Kondo of the University of Hawaii at Manoa.)

ANSWER: To review briefly, *item facility* is typically defined as the proportion of students who answered a particular item correctly. Thus, if 45 out of 50 students answered a particular item correctly, the item facility for that item would be 45/50 = .90, meaning that 90% of the students answered the item correctly and it is very easy. All the other classical theory estimates for norm-referenced and criterion-referenced item analysis techniques that I give in my book are based on that notion of *item discrimination*: item discrimination is defined as the item facility on the particular item for the upper group (usually the top 33% or so, based on their total test scores) minus the item facility for the lower group (usually the lower 33% or so); the *difference index* is defined as the item facility on the particular item for the past test minus the item facility for that item on the pretest; the *B-index* is defined as the item facility on the particular item for the students who passed the test minus the item facility for the students who failed. Calculating these statistics is easy as long as the definition of item facility remains the proportion of students who answered a particular item correctly.

However, in some cases, especially in classroom criterion-referenced tests, teachers want to give partial credit for items (i.e., give 1, 2, or 3 points for a particular item depending on how well students answer it or give half a point for a partially correct answer and a full point for a fully correct answer). In that case, because the item is not clearly right or wrong for each student the definition of item facility given at the end of the previous paragraph does not work, and as a result, none of the other item statistics can be calculated either. In such cases of partial credit, a somewhat different definition of item facility could be used: item facility is the average proportion of correctness for a particular item.

The trick to calculating this version of item facility is to put each student's answer to each question on a proportion score scale from 0 to 1. The scoring of each item on the actual test can be on a 0 - 3 integer scale, a 1 - 5 scale, a 0 - ½ - 1 scale, or a 0 - 3 decimal scale, but for purposes of item analysis, they all need to be converted to a 0 to 1 scale by dividing each student's item score by the total possible for that item. Thus a 0 - 3 integer scale (i.e., a 0, 1, 2, 3 scale) could be converted to .00 for a student who answered completely wrong (0/3 = .00), .33 for a student who got 1 point and answered one-third correctly (⅓ = .33), .67 for a student who got 2 points and answered two-thirds correctly (⅔=.67), and 1.00 for a student who got all 3 points and answered completely correctly (3/3 = 1.00).

The same principle could be applied to a 1- 5 scale, with scores of .00, .20, .40, .60, .80, and 1.00 being possible. For a 0 - ½ - 1 scale, the points on the scale could be converted to .00, .50, and 1.00. For a 0 - 3 decimal scale, the points on the scale of .00, .10, .20, …, 2.8, 2.9, and 3.0 could be converted to .00, .03, .07, …, .93, .97, and 1.0. And so forth.

Once each student's proportion score for each item is coded in this way, item facility is simply the average of all these values across students for each item. For instance, let's say that ten students took a four-item test scored with different weightings for each item (regular right/wrong for the first item, a 1 - 3 integer scale for the second question, a 1 - 5 scale for the third question, a 0 - ½ -1 scale for the fourth question, and a 0.0 to 3.0 decimal scale for the fifth question). The item results might look like the raw scores in Table 1.

Table 1. *Actual Scoring of the Example Test*

Student	Item I	Item 2	Item 3	Item 4	Item 5	Total Scores
Kimi	1	3	5	1	3.0	100
Sachiko	1	3	5	1	2.8	89
Keiko	1	2	4	½	2.1	85
Rieko	1	2	4	½	1.7	80
Mitsue	1	3	3	½	1.5	79
Hitoshi	1	2	3	½	1.0	70
Hide	0	1	2	½	0.9	64
Yoshi	0	1	2	0	0.7	50
Toshi	0	0	1	0	0.5	37
Hachiko	0	0	0	0	0.3	13

Table 2. *Proportion Score Equivalents for Each Item (and Item Statistics)*

Student	Item 1	Item 2	Item 3	Item 4	Item 5	Total Scores
Kimi	1.00	1.00	1.00	1.00	1.00	100
Sachiko	1.00	1.00	1.00	1.00	0.93	89
Keiko	1.00	0.67	O.80	0.50	0.70	85
Rieko	1.00	0.67	0.80	0.50	0.57	80
Mitsue	1.00	1.00	0.00	0.50	0.50	79
Hitoshi	1.00	0.67	0.00	0.50	0.33	70
Hide	0.00	0.33	0.40	0.50	0.30	64
Yoshi	0.00	0.33	0.40	0.00	0.23	50
Toshi	0.00	0.00	0.20	0.00	0.17	37
Hachiko	0.00	0.00	0.00	0.00	0.10	13
IF	0.60	0.57	0.58	0.45	0.4	
ID	1.00	0.78	0.73	0.83	0.71	

The results in Table 1 would be converted into proportion scores as described above for each person on each item as shown in Table 2. Then, with item facility redefined in this way, the item analysis could proceed with only slight variations from the usual classical theory calculations. Item facility (IF) becomes the average of the proportion scores. For instance, for Item 2 in Table 2, the calculations would be as follows:

$$IF = (1.00 + 1.00 + .67 + .67 + 1.00 + .67 + .33 + .33 + .00 + .00) / 10 = .57$$

Then, the item discrimination statistic (ID) could be based on the average proportion score for the upper group minus the average for the lower group. Where the upper group is defined as the top three students in Table 2 and the lower group is defined as the lower three students. ID for Item 2 would be calculated as follows:

$$IF^{upper} - IF_{lower} = (1.00 + 1.00 + .67) / 3 - (.33 + .00 + .00) / 3 = 2.67 / 3 - .33 / 3 = .89 - .11 = .78$$

Exactly the same principles could be applied to calculating the difference index (DI) and the B-index. Note that the results for IF, ID, DI, and B would be interpreted very much in the same way they are normally interpreted.

Other strategies exist for dealing with weighted items in item analysis[1]. However, the proportion score strategies that I explained in the body of this article seem to me to be the easiest to understand and carry out, I hope that you will find them practical and useful ways of dealing with item analysis for tests with weighted scores.

References

Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River. NJ: Prentice Hall Regents.

Brown. J. D. (translated into Japanese by M. Wada). (1999). *Gengo tesuto no kisochishiki*. [Basic knowledge of language testing]. Tokyo: Taishukan Shoten.

[1] For instance, item facility could be calculated as a simple average of the weighted scores shown in Table 1. In such a case, the values would simply be reported and interpreted relative to the possible values in the scale. For instance, the average for Item 2 in Table 1 would be 17/10 = 1.7, which could then be compared to the total possible for that item of 3 to determine whether or not it was difficult. However, using this method, the interpretation would be different for each type of item weighting, which could prove confusing. An alternative strategy for calculating item discrimination for weighted items would be to use computer power to calculate whatever correlation.

HTML: http://www.jalt.org/test/bro_6.htm **PDF**: http://www.jalt.org/test/PDF/Brown6.pdf