Item point-biserial discrimination

Luc Le - 13 June 2012

One common index used in item traditional statistics to assess item quality is item discrimination. This quantity refers to the degree to which an item differentiates correctly among examinees in the behaviour that the test is designed to measure. When the test as a whole is to be evaluated by means of criterion-related validation, the items themselves may be evaluated and selected on the basis of their relationship to a criterion. The typical classical index of item discrimination is some correlation measure between examinee performance on the item and their performance on the test (criterion). Positive values are desirable and indicate that the item is good at differentiating between high ability and low ability examinees.

Item discrimination and fit statistics to the Rasch model

Item discrimination is not parameterised in the Rasch model (and its extended models) but the model specifies that item discrimination or ICC slope in a test be held constant (Wu & Adams, 2007). The common item fit statistics, named as infit and outfit, developed by Rasch (1960/1980) and Wright (1977) can test the violations of the model assumptions of no guessing and homogeneity of item discrimination. In the dichotomous case, Margaret Wu has proven that these statistics are measures of the heterogeneity of item discrimination (Wu & Adams, 2007, *pp* 85-87). Varying the item discrimination will simulate misfit in the Rasch model (Baker, 1992). The variation of the item discrimination in a test and these fit statistics are asymmetric, as mentioned by Wright (1992) that "The relation between point-biserial correlation discrimination estimates (*rpbis*) and Rasch fit statistics (RFS) is nearly monotonic apart from the effect of item-person targeting on point-biserial ceilings" (p. 174).

Item point-biserial discrimination

When the examinee performance on the item is scored as 1 (correct) or zero (incorrect), the Pearson product-moment correlation between the item score (1/0) and their total score on the test is called the point-biserial correlation (r_{pb}).

Basically, the point-biserial correlation relates the examinee's item scores with their total RAW scores on the test. A slightly more robust version of this, called the corrected point-biserial correlation, calculates the relationship between the item score and the test score after removing the item score from the total test score. This is an appropriate correction because total scores that have the item in question embedded within them will have a spuriously higher relationship than total scores made up of only the other items in the test. This correction is particularly important for short tests where one item can dramatically impact the total score. But its impact diminishes as test length increases.

Additionally, when the item is multiple choice in format, it is also important to study responses to its distractors. The first step in calculating the point-biserial correlation coefficient for a distractor is to assign dummy values to the item. Traditionally, one is for any response to distractor zero is for any response to other item categories. Typically, for most measurement purposes, an item which is acceptable should have a positive point-biserial discrimination for the correct answer and negative point-biserial discrimination for each of its distractors or wrong answers (Millman & Green, 1989).

Formulae for item point-biserial discrimination

Let *Y* be a dichotomous variable (i.e, an item scored as 0 and 1) and *X* be one continuously measured variable (i.e., test score, include or exclude the examined item). Divide the data set into two groups, where *Y* is 1 or 0, respectively. Denote M_1 and M_0 the mean on the criterion variable X of the examinees, and n_1 and n_0 the number of examinees in the two groups respectively. Then, the mathematical formulae for the item point-biserial coefficient is:

$$r_{pb} = \frac{M_1 - M_0}{S_n}\sqrt{\frac{n_1 n_0}{n^2}}, \text{ or} \qquad (1)$$

$$r_{pb} = \frac{M_1 - M_0}{S_n}\sqrt{pq}, \qquad (2)$$

where $n = n_1 + n_0$ is the total number of examinees, S_n is the standard deviation on the criterion of all examinees, and *p* and *q* are the proportion of examinees who answer correctly or incorrectly to the item (*p* + *q*=1).

$$S_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (3)$$

We can test the null hypothesis that the correlation is zero in the population. A one-tailed t-test independent means is:

$$t = \frac{r_{pb}\sqrt{n-2}}{\sqrt{1-r_{pb}}} \qquad (4)$$

where *n*-2 is the degrees of freedom.

It can be seen from formulas (1) or (2) that r_{pb} is proportional to the mean difference of the two groups on the criterion variable X. Moreover, this quantity is likely to have a large magnitude when *p* is close to 50% or the standard deviation on the criterion of all examinees, S_n, is small. On the other hand, this quantity is likely to get a small magnitude when *p* is very big or very small, or S_n is big.

Command and outputs

ConQuest 3 is able to perform a traditional item analysis for all of the generalised items, including item discrimination index. Specifically, it can provide the item point—biserial (discrimination) as a correlation between the item scores and the test scores, formed by all items in the test (including the study item) or formed by all of the rest items in the test (exclusive the item).

Some examples of possible data layouts are as follows:

Example 1:

```
itanal format=summary >> itanal.out;
```

In this case, the summary format provides a line of information for each generalised item. The information given is restricted to the item label, facility, discrimination, fit and item parameter estimates. Particularly, column four (titled “Item-Rst Cor”) includes the corrected item point-biserial, that is the correlation between the item score and the total scores by all of the rest items in the test (exclusive the item). While column five (titled “Item-Tot Cor”) includes the basic item point-biserial correlation, where the test score is formed by all items in the test (including the study item).

Example 2:

```
itanal >> itanal.out;
```

In this case, ConQuest 3 performs a traditional item analysis for all of the generalised items and writes the results to the file itanal.out.

References

Baker, F. B. (1992). *Item Response Theory: Parameter Estimation Technique*. New York: Marcel Dekker, Inc.

Millman, J., & Green, J. (1989). The specification and development of tests of achievement and ability. In R. L. Linn (Ed.), *Educational measurement*, 3rd Ed., (pp. 335-366). New York: American Council on Education and Macmillan

Rasch, G. (1960/1980). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danish Institute for Education Research, 1960. (expended edition, Chicago: The University of Chicago Press, 1980.)

Wright, B. D. (1977). Solving measurement problems with the Rasch Model. *Journal of Educational Measurement*, *14*, 97–116.

Wright, B. D. (1992). Point-biserial correlations and item fits. *Rasch Measurement Transactions*, *5*(4), 174–174.

Wu, M. & Adams, R. (2007). Applying the Rasch model to psycho-social measurement: A practical approach. Educational Measurement Solutions, Melbourne. Available from: http://edmeasurement.com.au/_docs/RaschMeasurement_Complete.pdf