Metric Quality of Tests

The evaluation of learning is done by using tests and questions (i.e. "items"). The results obtained after administration of a test will be used to take various decisions. These may promote a student, select for admission to a school or employment sector; the results can also help to plan school intervention programs. We could list many other uses of the results obtained following an evaluation of learning. Also, it is necessary that the instruments used to gather information respect high standard of quality.

There are statistical analysis techniques to assess the metric qualities of a test. We will make the presentation of a few. In addition, we recall some concepts related to the presented techniques. We will address the items analysis in relation to their metric quality and no bias penalizing certain subgroups of evaluated students.

A. Items Analysis

It is important to consider the qualities of a measuring instrument and items that are used. It is the statistical process that is usually mentioned under the term items analysis. Several techniques can be used to ensure the medic quality of a test. We briefly describe what is meant by *classical item analysis* and techniques based on the *theory of item responses*.

1. Classical items analysis

The *classic items analysis* is associated to multiple choice answers or ones on continuous scale or partial credit. It allows estimating the metric quality of each item belonging to a test. This type of analysis enables the estimation of the metric quality for each item that is part of a test. Three sets of statistical characteristics are usually retained.

- *Difficulty* – For each item included in a test, we must calculate the relative difficulty; a test must contain easy questions, moderately easy, more difficult ones to collect information on all students being evaluated. For multiple-choice questions, it is the proportion of students who correctly answered the question. For items associated to partial credits, we calculate an average result.

- *Discrimination* – The questions asked to students should allow to clearly identify those who have mastered what is measured and those who have not mastered the assessed concepts, skills; we need to know certain characteristics of these two groups of students. Statistical indexes provide information on the ability of items to clearly identify the students called "successful" and those less powerful.
- *Consistency* – As we know, a test consists of several items or questions. We may want to know the consistency that should be established between the various collected information. This is the internal consistency of the instrument.

EduStat software allows for such analysis. Thus, for each multiple choice item of a test, a report shows the number of individuals, the proportion of those who succeed the item, the coefficient of item / test correlation. It is also possible, in parallel, to obtain other statistics (e.g., success percentages of those who answer correctly the item and those who failed, the Alpha coefficient recalculated excluding the examined item, statistical discrimination according to each choice of response, the grouping of items relative to the specification table, the description of each item). Moreover, for each item witch the answers are on, a continuous scale (partial credits), the report provides the number of individuals, the minimum and maximum scores observed, the average obtained, the expression of that average on a scale 0 - 100 and the correlation coefficient item / test. For all of the test and compared with each grouping of items, the report provides internal consistency coefficient in the form of Cronbach's Alpha, the standard error of measurement and standard deviation.

2. Items analysis using “Item response theory”

The item response theory is a set of statistical techniques whose aim is to assess the quality of a measuring instrument and items that make it up. It is possible to retain models taking into account one or the other next parameter (or all three):

- *Discrimination*, that is to say, the item power to properly identify students who have mastered what is measured and those who have not mastered the concepts, skills or competencies assessed.
- The item *difficulty*, that is to say, the more or less easy item for students included in the test.
- The *pseudo luck,* that is to say the possibility for the student to obtain a correct answer at random without actually possess the required knowledge.

There are specialized software to perform these calculations. EduStat allows the preparation of files that can be used by such software. Furthermore, the EduStat software can use the results calculated by other software (for example, XCalibre) for tracing curves illustrating the results obtained to the analysis performed for each item of an event.

B. Analysis of Differential Item Functioning (DIF)

It is important to ensure that some issues part of an evaluation instrumentation do not favor a subgroup of students at the detriment of others. This is the way of examination that may affect several socio-economic and cultural dimensions. We may want to consider if there is presence of an angle to two subgroups of students (for example, boys and girls) or more (e.g., administrative regions). The analysis techniques will be different for each of these situations. The items identified as "biased" from the value of the indices calculated should be subtracted from the test or modified.

1. The "omnibus" analysis

The technique called "Omnibus" allows the calculation of indices in relation to several groups simultaneously. The produced report contains a set of statistics for each item selected for analysis. In addition, a graph illustrates the suitability of the theoretical distribution to empirical distribution that evaluates the coordinates of rank values (axis "X") and DR values (axis "Y"). A good fit between the values "rank" and "DR" is obtained by a graph in which the points (x, y) draw a straight line. Any point deviating significantly from this straight line is interpreted as a potentially marginalized item.

2. The analysis using two groups

If we have to examine the differential item functioning comparing both groups, Mantel-Haenszel technique can be used. The report produced by EduStat software as a result of the use of options for calculating the coefficients "M-H" contains the following:

- identification of reference groups and compared one;
- for each item, the indexes Alpha, Delta and standard error attached to the Delta;
- for each item, the confidence interval of Delta;
- classification of items with respect to three categories of bias:

  - C: priority item to consider with respect to the possibility a differentiated item functioning;
  - B: second category of items likely to have a differentiated functioning;
  - A: item that is not identified as likely to have differentiated functioning.

Here are some notes for interpreting the statistics produced in the differential functioning of the item analysis by the Mantel-Haenszel method.

> The items identified by the label "C" are those that may have biases affecting their functioning. The absolute value of Delta is equal or greater than 1.5 and is significantly different from 1.
>
> As for the least likely to be biased items ("A" category), the absolute value of Delta is less than 1 or is not significantly different from "0".
>
> Other items whose absolute value is between 1 and 1.5 are found in the "B" category and may be slightly biased.

A Delta positive value indicates that the item was harder for students belonging to the reference group. A negative value therefore identifies a most difficult item for the group compared to the reference group.