

# Testing the Test

drs. Peter de Jong  
University of Groningen  
Chemistry department  
Nijenborgh 4  
9747 AG Groningen  
The Netherlands

## ***Introduction***

As in many educational situations, we at the University of Groningen want to know if the students have learned something during the course. Usually the students take an exam at the end of the course. The students who pass the exam are happy and the ones who fail, are sad and have to take the exam again a few months later.

During the course there are no tests, so the students don't know if they really understand the subjects. And the teacher doesn't know if the subjects he teaches are clear.

As a solution for this problem, one can make some tests, which are made during the course. The problem with this system is the lack of time the teachers have. They don't have time to correct those extra tests and give feedback to the students.

That's why we wanted to make a computer program for testing. This is fairly easy done, when using multiple-choice (MC) questions in Toolbook Instructor. To keep the program suitable for different courses, the questions are put into a database. From that database a random selection of questions is taken. The statistical analysis is also valid for a test with a fixed number of questions.

Because of the lack of time the professor can't make those questions himself, so his assistants have to make them. The assistants are researchers and generally they have little knowledge of didactics. As a result not all the questions are as valid as they should be. Of course you can give some instructions to the assistants how to make MC questions, but still there will be questions that don't test what the teacher wants to test.

Because we want the assistants to enter the questions in the database themselves, there is no check of the validity of the questions. We decided to put some extra fields in the database for some variables so there is a way to check the validity of questions.

## **Qualitative criteria of MC questions**

We use MC questions with four possible answers. A question with the answers is called an item. An item has to meet the following qualitative criteria:

- **Relevance or significance**  
The test has to test what you want to achieve with the instruction.
- **Balance**  
The items have to test all the subjects taught during the instruction.
- **Objectivity**  
Are the correct answers really correct? Experts in the subject have to agree with each other, about the correctness of one alternative answer and the incorrectness of the other alternative answers.
- **Differentiation and discrimination**  
The item has to distinguish two groups of students:
  - a. Those who have studied the subject sufficiently, or who understand the subject sufficiently.
  - b. Those who have not studied the subject, or who don't understand the subject.

The correct answer has to be recognised easy by the students of group *a*, while the other alternatives have to be equal attractive to the students of group *b*, because the alternatives seems to be logically, or because the alternative represents a frequently occurring mistake.

- **Specificity**  
The change of answering the item correct, without studying the subject, i.e. with common sense or logical thinking, has to be small.
- **Fairness**  
The students have to know the answer or can deduce it, in view of the given instruction.
- **Appearance and formulation**  
The question has to be as short as possible, clear and accurate. Try to avoid negative or absolute phrases (not, never, always, etc.)

With these criteria in mind, it's possible to make valid items. But there's still a chance to misjudge a question and/or an answer.

That's why items have to be reviewed after the test has been taken. Next you can correct the less valid items and use them next time again or you can decide to remove that item. For reviewing the items statistical data obtained by the test can be used. This data consists of the following values, which will be described below: the p-value, the a-value and the item-test-correlation coefficient.

## ***p- and a-values***

The p-value is the relative frequency of the correct alternative. It's the quotient of the number of students that made the item correct (P) and the number of students that made the test (N). The p-value can be between zero (nobody has chosen the correct answer) and one (everybody has chosen the correct answer).

$$p = \frac{P}{N}$$

The a-values are the relative frequencies of each of the incorrect alternatives.

### Interpretation of the p- and a-values

The p-value of an item should not be too low. A low p-value means that the item is too difficult, or that one of the other alternatives was too attractive. The latter will be shown with a high a-value on the alternative. When no one has studied the subject, a p-value of 0.25 is expected when using an item with four possible answers, because of the possibility to guess the correct answer.

The p-value may not be too high. A high p-value can mean that the item could be answered with common sense or logical thinking, without studying the subject. This is not what you want for a test, because you want your items to discriminate between the good and the less fortunate students.

So what p-value is desirable?

The only effect of items with a low or a high p-value is lowering or raising the total score of the test. Most times a test is meant to discriminate between the good and the less fortunate students. Therefore the p-value for items with four choices should be between the values 0.25 and 1.00, and the average (0.63) is desirable. Items with a p-value over 0.90 are too easy and items with a p-value lower than 0.45 are too difficult and most times obtained by guessing.

Low p-values can be acceptable, if they are coupled with a high item-test-correlation (see next paragraph).

The a-values of an item have to be much lower than the p-value. If the a-value is lower than 0.05 the distracter didn't function. If an a-value is higher than or close to the p-value of an item, the item can be ambiguous except when there is a high item-test-correlation.

### **Checking for item-test-correlation**

Because the item score can have only two values (zero or one), the next formula can be used to determine the item-test-correlation coefficient ( $r_{it}$ ):

$$r_{it} = \frac{\bar{U}_i - \bar{Y}}{s_y} \sqrt{\frac{p}{1-p}}$$

$\bar{U}_i$ : The average score for the test of all persons having the item correct.

$\bar{Y}$ : The average score for the test.

$s_y$ : The standard deviation of the test score.

$p$ : The p - value of the item.

This coefficient can vary from  $-1$  to  $+1$ . When it's  $+1$ , the item correlates perfectly with the test, i.e. the persons who scored the item made the test well.

When  $r_{it}$  is  $-1$ , the item score correlates inversely with the test score, i.e. the persons who scored the item, failed the test.

When  $r_{it}$  is zero, there is no correlation in test score and item score.

The following appreciation of an item can be given on account of the value of  $r_{it}$ :

$r_{it}$	appreciation
0.40 or more	very good
0.30 – 0.39	pretty good, can possibly be improved
0.20 – 0.29	has to be improved
less than 0.19	bad, remove or revise totally

### **The reliability of the indices**

The p- and a-values have a reliability of 80%, when 25 students take the test. So the test group for the questions doesn't have to be large, to have quite reliable p- and a-values.

The item-test-correlation is less reliable. A group of about 100 persons is needed to obtain a reliability of 50%. This is the reason why items can't be judged only on these indices.

### **Which extra fields do we need in our database?**

To keep track of the items' validity we need some fields in our database to save the parameters for p- and a-values and  $r_{it}$ .

For the p-value P and N are needed. For the a-values N is also needed and for every alternative an extra field is needed to keep track of the number of persons that have chosen the alternative.

Finding the parameters for  $r_{it}$  is a little harder; these parameters are compounded. Let's look at them one by one.

$$\bar{Y}_i = \frac{\sum \text{total scores of persons having the item correct}}{P}$$

For the numerator an extra field is needed.

$$\bar{Y} = \frac{\sum \text{all scores on the test}}{N}$$

For the numerator in this formula an extra field is needed.

$$s_y = \frac{1}{N} \sqrt{N \sum (\text{all scores})^2 - (\sum \text{all scores})^2}$$

For 'all scores squared' a field is needed. The other parameters are already known from previous formulas.

All together there are eight extra fields needed in the database to allow a check for item validity.

### **Conclusion**

With eight extra fields in the database it's possible to hold all the parameters to check the items for their validity. Of course the teacher still needs to check the items, but with the p-value and the item-correlation-coefficient it's easier to review the questions and their alternatives quickly.