Correcting Scores of Tests Taking into Account the Guessing Factor: Yes or No?

Jorge Valadares
Universidade Aberta e Unidade de Investigação e Desenvolvimento em Educação

Abstract

Some time ago, in my Department, some colleagues put me the following question: when we classify the objective tests, is it necessary to introduce a correction factor, having in account that students can get correct answers by chance? If yes, how to proceed? I studied and thought about the subject, and this paper synthesizes the answer that I presented in the Department.

1. Introduction

The problem that arises is the following:

Students perform tests of various formats. They respond to certain types of items and get right or not. They respond to other types of items and have more or less success. They leave more or less items unanswered. It is an obvious fact that some answers may be correct because the students have simply guessed the answer. If so, is or not legitimate to introduce a correction factor in the final scores of each student? If yes, how to proceed?

2. We should take into account the type of items

As we know, there are two major types of items: items where students have to construct the answer and items where students have to select the answer.

Here's an example of an item of the first kind, "Who discovered Brazil?"

In this section, the sample space, including all students that do not know anything about the discoveries and much less about the discovery of Brazil, is the set of all possible names. It is so large that the probability of getting the correct answer by chance is practically zero. With such items, more or less difficult no matter, **the correction for guessing is unfounded**. However, this can no longer happen with items of response selection as is the case, specifically, of the multiple-choice items.

If, for example, a multiple-choice item has four options, with the general form

"STEM OF ITEM":

A. 1st Option
B. 2nd Option
C. 3rd Option
D. 4th Option

the sample space is

E = {A B, C, D}

so the probability, P , of a student getting right by pure guessing (only by luck) is

$$P=\frac{\text{number of favorable cases}}{\text{number of possible cases}}=\frac{1}{4}=0{,}250=25\%$$

as the four options are equiprobable for him.

Well, but a multiple-choice test is certainly not just composed of an item of this type. We need to use tests with several items. At this stage of our reflection, it is important to remember two important aspects of assessment.

The first aspect is that the assessment is highly contextual, depending on its type, its purposes and the conditions under which it occurs. It follows that the way we build, classify and express the results of a test depend on its purpose and should never be extrapolated these results to other contexts. Another important aspect is that assessment is destined to formulate well-grounded value claims, based on referents previously adopted, in order to make good decisions, taking into account the purposes of the test. But these referents differ and these differences will have implications in the conclusions, as we shall see in the next section. Finally we must have in account that there are many differences among examinees in their tendencies to guess the answers and in the way they try to guess.

3. Types of tests

We now distinguish two types of tests, with quite different purposes.

3.1. The test is intended to assess the proficiency level of certain objectives

In this case we construct a test of reference clearly criterial. Three important characteristics of a criterion-referenced test are the following ones:

a) The domain of objectives to assess is more or less narrow, with several items used to measure the so called level of proficiency of each objective.
b) What we intend with the analysis of the answers is to describe the proficiency level in each objective [1].
c) The results must be expressed in terms of the proficiency level of the test objectives. The following table illustrates, very clearly, the meaning of the proficiency level:

Table 1. The proficiency level

PROFICIENCY IN WHAT? / PROFICIENCY FOR WHOM?	IN AN OBJECTIVE	IN THE SET OF OBJECTIVES
OF ONE STUDENT	He/she gets of a bar graph two of the three asked information	He/she demonstrates proficiency in six of the ten objectives assessed with the test
OFA GROUP OF STUDENTS	90% of students reach the objective regarding the reading of bar graphs	80% of the students satisfy in the test

Some proficiency levels depend on the others, as the arrows try to show. For instance, the level of each student's proficiency in each objective determines his/her proficiency level in all objectives assessed with the test.

The sub problem that arises now is this:

If we built a given number N of questions for the same objective, what is the minimum proficiency level that we should demand to guarantee that the objective was sufficiently reached by a student, having in account that he/she might have gotten right by chance?

This sub problem leads us back to the question of guessing.

Let us suppose that we used two questions to assess a given objective (N = 2), each of them with 4 options (n = 4). A student responded correctly to both items. But he could have been lucky enough to set right by chance in the two items without knowing the correct answers. Another student responded correctly to one of the items. But he could have chosen by pure guessing the correct option and not have been lucky enough to set right by chance in the other option, as he/she did not know the correct answers. Or he/she responded correctly to one of the items because he knew perfectly the answer. In terms of guessing everything is possible, otherwise no one got right in the most several lucky games, and this fact takes us to the field of the probabilities.

What is the probability of setting right at random in 1 question of the two questions with 4 options, constructed to assess the same objective?

The probability of making the right choice in the first question is 1/4. The probability of wandering in the second question is 3/4. Then, the probability of the united production of these two events, that is, to choose the right option in the first question and wrong option in the second one, is the product of the respective probabilities:

$$\frac{1}{4}\times\frac{3}{4}=\frac{3}{16}$$

In a similar way, we conclude that the probability of choosing a wrong option in the first question and a right option in the second one is the product of the respective probabilities:

$$\frac{3}{4}\times\frac{1}{4}=\frac{3}{16}\ .$$

Then, as these two events, to get right in any one of the questions and to wander in the other, are independent, the probability of one of them, whichever it is, is just the sum of the probabilities of each of them:

$$\frac{3}{16}+\frac{3}{16}=\frac{6}{16}=0{,}375=37{,}5\%$$

In the same way, the probability of getting right, by chance, in both 2 questions with 4 options, is:

$$\frac{1}{4}\times\frac{1}{4}=\frac{1}{16}=0{,}063=6{,}3\%$$

These probabilities of getting right by chance in all of the questions, or in half of them, for instance, decrease with the number of options. We saw that the probability of setting right at random in 1 question of the two questions with 4 options is 37,5%. And the probability of setting right at random in both 2 questions is 6,3 %. If, for example, the teacher decides to construct two questions with five response alternatives, instead of four, with a completely analogous reasoning we conclude that:

- The probability of getting right, in 1 item (any one of two items) with five options, answered completely at random, is

$$\frac{1}{5}\times\frac{4}{5}+\frac{4}{5}\times\frac{1}{5}=\frac{8}{25}=0{,}320=32{,}0\%$$

In turn, the probability of getting right, in 2 items with five options, answered completely at random, is

$$\frac{1}{5}\times\frac{1}{5}=\frac{1}{25}=0{,}040=4{,}0\%$$

Sometimes is not easy to design items with 5 options to assess an objective, but we can use 4, 5 or more items with 4 options instead of two. The probability of getting right by chance in all the questions, or in an half of them, or in any percentage of them, decreases with the total number of questions.

For, example, we saw that the probability of getting right, by chance, in 1 of 2 questions (50 %) with 4 options, is:

$$\frac{1}{4}\times\frac{3}{4}+\frac{3}{4}\times\frac{1}{4}=\frac{6}{16}=0{,}375=37{,}5\%$$

What is the probability of getting right, by chance, in 2 questions (whatever they are) of 4 questions (50 %) with 4 options?

The probability of getting right by chance in the first two of 4 items with 4 options, wandering in the other two, is:

$$\frac{1}{4}\times\frac{1}{4}\times\frac{3}{4}\times\frac{3}{4}=\left(\frac{1}{4}\right)^2\times\left(\frac{3}{4}\right)^2=\frac{9}{256}=0{,}0351=3{,}51\%$$

But they can also respond correctly to the first and the third questions, or they can respond correctly to the first and the fourth, or to the second and the third, or to the second and the fourth, or to the third and the fourth. And the probability is the same, 3,5% for all happenings. Then the probability of getting right in 2 questions, whatever they are, is:

$$6 \times 3{,}51\% = 21{,}1\ \%$$

Another example: what is the probability of setting right by chance in 3 of 5 questions with 4 options, constructed to assess the same objective?

The probability of getting right by chance in the first three of 5 items with 4 options, wandering in the other two, is:

$$\frac{1}{4}\times\frac{1}{4}\times\frac{1}{4}\times\frac{3}{4}\times\frac{3}{4}=\left(\frac{1}{4}\right)^3\times\left(\frac{3}{4}\right)^2$$

Recurring to the combinatory analysis we know that there are different ways of distributing the 3 successes for the 5 items, then the probability of getting right by chance in 3 items, whichever they are, is

$$^5C_3 \times (1/4)3 \times (3/4)2 = 0{,}088 = 8{,}8\%$$

In general, we can calculate the probability of setting right by chance in n of N questions with equiprobable answers, constructed to assess the same objective with the formula

$$^NC_n\times\left(\frac{1}{a}\right)^n\times\left(\frac{a-1}{a}\right)^{N-n}$$

With this formula we can obtain the probabilities in the Table 2.

In this Table, the percentages of setting right by chance in half of all the items are particularly signalized. We can see that with 6 items from 4 options designed for a given objective, the student has a probability of 13.2% of getting right in half of them, 3 items.

Table 2. Probabilities (in percentages) of getting right, by chance, in questions with 4 options

	Number of items								
Number of right answers	1	2	3	4	5	6	8	10	12
1	25,0	**37,5**	42,2	42,2	39,6	35,6	26,7		
2		6,3	14,1	**21,1**	26,4	29,7	31,1		
3			1,6	4,7	8,8	**13,2**	20,8		
4				0,4	1,5	3,3	**8,6**		
5					0,1	0,4	2,3	**5,8**	
6						0,0	0,4		**4,0**
7							0,0		

To have success only in 3 items of a group with 6 items designed to assess an objective may be or not considered a poor result in relation to the proficiency level in that objective, because the student may have tried to guess all the answers by guessing and has 13,2% of probability of getting right in 3 items. Nothing impedes that we stipulate the success in 4 of the 6 questions as a cut score to reveal enough proficiency in that objective, as the probability of getting right by chance decreases to 3,3%. Or we can use items with 5 options, and it is unlikely to guess the same four answers completely at random, as the chance for that is only 1,5 %.

Table 3. Probabilities (in percentages) of getting right, by chance, in questions with 5 options

	Number of items							
Number of right answers	1	2	3	4	5	6	8	10
1	20,0	**32,0**	38,4	40,9	40,9	39,3		
2		4,0	9,6	**15,4**	20,5	24,6		
3			0,8	2,6	5,1	**8,2**		
4				0,2	0,6	1,5	**4,6**	
5					0,0	0,2		**2,6**
6						0,0		

The question that arises now is:

If a student has 3 right choices in 5 multiple-choice questions, how many items have been answered at random? None, one, two, three, four or five? Only he/she knows. So what criteria to adopt? Let us consider that a student achieved the objective if has four or five correct responses, and did not

achieve the objective with 3, 2, 1 or none correct answers? Or we must try to quantify the level of proficiency subtracting some amount from the 100% of proficiency, depending on the number of wrong items (the same amount for each wrong answer or not)?

Subsequently, we will go to face two possible processes to meet a classification criterion. Each of them is included in the assessment literature and has already been adopted many times. Now we go to pass to another kind of tests.

3.2. The test is intended to assess the general knowledge of students

In this case, the test is intended to assess the summative degree of dominance of a more or less extensive knowledge field. As is common in summative assessment, the tests cannot be criterion-referenced tests (several items to assess each objective) but rather must be norm-referenced tests. Its main features are as follows [2]:

- The domains of objectives and contents to be evaluated is more or less broad.
- The questions must match with the broad conceptual framework that we intend evaluate in a general perspective.
- The major emphasis is in comparing students, trying to discover individual knowledge differences and no concern in mastering this or that objective bases in this or that criterion.

An only test designed to assess competences or proficiency levels in learning objectives in a whole course and that only evaluates a goal here, another there, is a bad test because it lacks content validity [3], [4]. However, in this case, what is intended by the analysis of responses is to compare the overall achievement of each student with the overall achievement of others who were submitted to the same type of assessment, under identical conditions. In some countries there are tests named standardized achievement tests designed to be as marketable as possible, across districts, great regions or the nations as a whole. They include norms, previously built, which are tables of results obtained with these tests. These results, many times is percentiles, have been obtained with representative samples of the students population to be assessed with the standardized tests. In the ordinary case of our schools, these tests do not exist but somehow we can consider as a norm, when we use a normative test, the overall results of all the students that were evaluated.

Having in account this important difference between criterion-referenced testing and norm-referenced test, let us return to the problem of whether we must use a correction for guessing or not, and, if yes, try to meet a classification criterion.

3.2.1. 1st Process: to use a mathematical expression that appears in some assessment manuals. Let us suppose that a student was submitted to a criterion-reference test, designed to evaluate if he dominates a determinate objective. To assess that objective, he answered to N questions, each one of them with n options. Let us also suppose that the student answered C + E questions, where C is the number of right answers and E the number of wrong answers So, we have as variables:

N = number of questions
n = number of options for each question
C = number of right answers
E = number of wrong answers

The number E of wrong answers may vary between 0 and N - C. The extreme value 0 corresponds to the situation where the student did not respond to more questions than those who got right and the extreme value N - C corresponds to the situation where the student answered all test questions.

Some authors argue that to offset the guessing factor, we must give a score given by the following formulas:

$$S = C - \frac{E}{n-1} \quad \text{se} \quad C \geq \frac{N}{n}$$

$$S = 0 \quad \text{se} \quad C < \frac{N}{n}$$

This expression was based on the following principle: a student, answering a test of N questions with n options, has the “whole chance" of guessing the answers in C = N/n of those questions.

If a student got right in those N/n questions, wandering in the remaining E = N - N/n questions, she/he did not go besides the number of questions proportionate by the chance of getting right; then, her/his proficiency level must be zero: S = 0.

Based on this principle, we deduce easily the expression above to allow obtaining the note S = 0 when the number of right questions is N / n and the number of wrong questions is N - N / n. If we subtract the number of wrong subjects, N -N/n, to the number of right subjects, N/n, we don't obtain zero. Nevertheless, if we divide the number of the wrong questions for a factor, x, in order that the difference became zero, a simple equation shows that the value of x is n- 1, in agreement with the given expression.

Exemplifying for a group of 6 questions for an objective, each one with 4 options, if the student gets right in the 6 questions (C = 6, E = 0), his/her proficiency level will be S = 6 – 0/3= 6 points, the maximum level, what is fair. If solves only 4 questions, getting right in all of them (C = 4, E = 0), his/her proficiency level will be S = 4 - 0/3 = 4 points. It is not introduced a correction factor, as the student did not try anything wrong, just not responded. If he/she answers to all the 6 questions getting right in only 3 of them (C = 3, E = 3), his/her

proficiency level will be S = 3 - 3/3 = 2 points. In this case the student score is corrected with a discount of 1 point.

The great critic to this process is that it is based in a principle without legitimacy. The student has not “the whole chance" (probability 100%) of getting right by chance in N/n questions, as one can see with the Table 2 and Table 3. The correction factor given by the above presented expression is exaggerated, when compared with the student's probability to get right by guessing. Then, for instance, in a group of 10 items with 4 options, if a student got right in half of the items, we have:

- Probability of getting right in five items wandering the other five: 5,8 %
- Discount foreseen by the previous expression: 5/3 × 1/5 = 1/3 = 33,3 %.

3.2.2. 2nd Process: to base the correction factor in the tables of probabilities of getting right by chance. We continue considering the criterion-referenced testing case.

Let us suppose that to assess a given objective the teacher prepared 4 multiple choice questions, each of them with 5 options. The table 3 shows us that the probability of getting right at random is:

- in 2 questions, 15,4 %;
- in 3 questions, 2,6 %;
- in all the 4 questions, 0,2 %.

As the purpose of the group of questions is to assess the proficiency level of a fixed objective, it is important to have in account what kind of objective it is.

First, we go to consider that it is an objective of restricted extent, concerning a quite structured situation, so that the universe of answers is limited and totally attainable by all students. Some authors designate it an objective at the minimum essential level. If, for instance, such objective is a prerequisite of future learning, it will be of demanding a quite high attainment level by all students. In this objective, each student must reveal a high level of performance. Having in account the student's relatively high probability to have 2 correct answers by chance, in 4 assessment questions to the same objective, the teacher can demand 3 right answers, as minimum proficiency level, since the probability of having 3 right answers by chance is only 2,6%. And, needing to attribute a score (many times these objectives are just qualitatively appraised, as it is only important to know whether the objective was reached or not), the teacher would give the score attributed to the objective to the students that had answered correctly to four or three questions, and zero to the others, considering that two or less correct answers do not guarantee that the students reached the objective; they might have tried to guess total or partially the correct answers.

Another different case is when the objective is of broad extent, it concerns a situation very little structured, where the universe of «good» answers, that prove that the objective was totally reached, is limitless and indefinite. Some authors call it a development objective. In this case, it is natural to wait the most varied attainments of the students, depending on the cognitive state in the course for reaching that objective. Now the teacher can vary the difficulty of the questions used to assess the same objective, to discover how far the students get to arrive. If the teacher has to classify the students quantitatively, he/she can give the maximum score to the students who got right in four items, and to give scores as minors as more items have been failed.

In both cases, any correction factor, similarly to what happens to the correction based on the formula, take for granted some presuppositions that may not be verified in many cases: that the tested student chose the options "blindly", that is, all possible answers have the same probability; that all the incorrect answers resulted from a choice based exclusively on luck and not on a lapse, or on a reading mistake, for instance, or even based on an ambiguity of the question statement, as unhappily happens more times than it would be desirable.

On the other hand, these corrections for guessing always end for putting the tested students in unequal conditions in what respects to the guessing "technique" and its absence. From students that first eliminate less plausible options and, only then, try the guessing, to the honest students that don't try to guess, there are all the possible situations. We must have in account that the need to put several items for each objective implies that the number of assessed objectives never will be very big, unless the students have a long time to accomplish the test, what rarely happens.

The results of a criterion-referenced test will have a very weak content validity if it is elaborated to assess a great matter extension, with lots of objectives, as it happens many times in the exams. In this case, which we intend to gain with the preciosity of the correction for guessing, not justified in many cases, we lost in what concern to reach one of the most important technical characteristics that the test results should have: reliability, necessary condition to validity.

If we construct a speed test, with many questions for each objective, and we intend to verify the proficiency level, based on more or less correct answers, it can happen that many students, not having any correction factor by random answers, bet seriously in the luck.

In this case, a good student, that honestly only answer the questions that is prepared to answer, can arrive at the end of the time with many less answered questions than the bad students that tried to guess. In this case, a serious reliability problem can appear and

the factor correction can be justified. As it is known, we should try to increase as much as possible the reliability of the test results, because this characteristic is a necessary condition, but no enough, of validity. But, it was already shown that the reliability of a test only increases with the introduction of the correction for chance by guess, if a positive correlation coefficient exists between the number of omitted answers and the obtained scores [3], [4]. This only happens when the tests are really very extensive, and, consequently, students with better scores are those that leave more questions without answering them, because the worst students answer many questions at random. A test of this nature should be calibrated so that the best students (that the teacher already knows) can answer practically to all or almost all items. The worst students could answer at random to many questions, but the theory of probabilities will be implacable for them.

Let us now consider the norm-referenced test case. In summative evaluation it is frequent to use multiple-choice tests with a large number of questions which objective is not to have a deep idea of the objectives that students dominate more or less, but just to know the evaluate individual differences as reliable as possible in what concerns to the performance of a students' group when compared with the performance of other people. The objective is not to compare the results with a somewhat arbitrary standard representing mastery of a content domain, as is vulgar, because many times the tests are far from being purely norm-referenced tests. Let us consider a summative test with 36 multiple-choice questions, each one with four options. Then we have:

$$N = 36 \text{ and } n = 4$$

If a student has 36 correct answers (C = 36, E = 0), the formula that is written before implies the score

$$S = 36 - 0 / 3 = 36 \text{ points}$$

That is the maximum score possible.

If a student solved only 27 questions and all of the choices are right (C = 27, E = 0), the formula implies the score:

$$S = 27 - 0 / 3 = 27 \text{ points}$$

The student score was not affected by any correction factor.

If a student solved all the test with only 27 correct answers (C = 27, E = 9), the guessing formula gives

$$S = 27 - 9 / 3 = 24 \text{ points}$$

The total score that results of correct answers was affected by a correction factor of 3 points.

If a student solved the entire test with just 9 correct answers (C = 9, E = 27), the formula gives the result:

$$S = 9 - 27 / 3 = 0 \text{ points}$$

Maybe the student considers himself penalized in this case, if he/she answered with conscience and knowledge and studied other subjects and only did not solve correctly more questions as he made some calculus errors or interpreted wrongly some questions not very clear. But if studying badly he was «imprudent» trying to select by chance many answers, is a fact that he only answered correctly to a quarter of the test questions, having each of them four options.

It was questioned the legitimacy of the principle which the correction formula is based on. So, the student of this last example has not «all the chance» of hitting in a quarter of all the answers and the students in general have not «all the chance» of hitting in N n correct choices, as we could observe calculating the probabilities as it is explained before.

In this kind of tests, the contents field and the assessment aim are broad in scope, and it is natural to expect the most varied attainments by students, depending on the cognitive state in which they are on their way to achieving that aim. With so many questions, the probabilities of getting right in significative proportions of questions by chance are very low, and it is not probable to have scores very affected by chance The teacher can now vary the difficulty of the questions that evaluate the same aim to see how far they can go to the students. In this case, norm-referenced tests having many questions with four or five response options, no correction must be make for a hypothetical guess, as the correction based on the formula is based in two assumptions that in many cases could not be verified:

1st - That the students chose the options tested 'blindly', that is, all options are equiprobable.

2nd - That all the incorrect answers were the result of a choice at random, and not a mistake in calculation, or a read error, for example, or even, as unfortunately happens more often than would be desirable, for an ambiguity of some questions.

Moreover, these corrections always end up putting students tested in inequality with regard to 'technical' guessing and the absence of guessing it. So:

a) There will be honest students that simply do not even try to guess what they clearly do not know.
b) Others will try to guess in a form completely 'blind'.
c) And finally, there are still others with experience and insight, which will eliminate less plausible options first, and only then, will deliver the guesswork among other options.

4. Conclusions

Concerning to criterion-referenced tests we saw that the correction for guessing, although does not have a very solid base, can be introduced if we adopt some criteria based on tables of probabilities. It will

function as a way of demobilizing the students' tendency for answering at random in the questions whose correct answers they are unable to reach. The less arbitrary criterion is based on the dichotomy he/she reached – she/he didn't reach the objective. It should be applied to the objectives at the minimum essential level.

The correct questions by chance could affect the value claims that we want to make about the student cognitive state in what refers to the evaluated objectives. But if we have a sufficiently big number of questions for each objective and, besides, the questions have 5 or more options, the introduction of any correction factor in the obtained score becomes perfectly unjustified, since the students have enough time for thinking about the questions and answer honestly to all of them. If we build a test with a minimum of 12 multiple-choice questions with 4 options or with a minimum of 10 questions with 5 options for each objective, it becomes perfectly unjustified any correction factor. The Table 2 shows that, when we have 12 questions with 4 options, the probability of getting right in half by total chance is only 4 %, and we can always require more than an half correct answers as the «cut score« to consider the attainment of the objective. On the other hand, the Table 3 shows that, when we have 10 questions with 5 options, the probability of getting right in half by total chance is only 2,6 %.

With norm-referenced tests the correction for guessing is yet more unjustified. We can construct tests with more than 20 items and the probabilities of getting significative proportions of right answers by chance are near zero.

When we construct a multiple-choice test, or a test including two subtests, one of them with multiple-choice items, and another with other kind of items, we must have in account the tables of probabilities and the need to use a sufficiently big number of questions, giving time enough to all of the well prepared students can answer to all (or almost all) questions, to avoid a reliability problem.

And, obviously, it is demanded that all the questions have the largest possible technical quality. Concerning to multiple choice questions, the technical quality demands that [2]:

a) the stem must presents a clearly defined problem (clarity doesn't mean simplicity or triviality);
b) the stem must presents as much information as possible, so that the options can be as short as possible, but, at the same time, with all the material needed to make the problem clear and specific;
c) the options must be as homogeneous as possible in what concerns to the content, the form and the extension;
d) the options must be written in a way grammatically coherent with the stem, and disposed in a natural order, if this exist;
e) the options “None of these” or “None of the above” must be used only when the key answer can be classified unequivocally as correct or incorrect;
f) the “distracters” should be plausible and attractive, no manifestly wrong for the students;
g) the questions must use novel material and the problems must assess understanding or ability to apply principles or theories;
h) the questions must use the negative only sparingly;
i) the questions must have one, and only one, correct or clearly best answer.

From all the above, it appears that the correction for guessing does not a very solid base to be conducted, even in these types of tests. However, if a test is highly formative, if what is sought is merely verify the level of achievement of some important objectives of minimum essential level, and we intend to demobilize the tendency of many students to guess on items that do not know respond, we can introduce a criterion based on probability tables. The less arbitrary is one that is based on the dichotomy achieved - not achieved the objective.

Using an enough number of well formulated multiple-choice questions (having in account the presented tables of probabilities), in the great majority of the cases we can abdicate of any correction for guessing. Using subtests to the same objectives, we can make studies of reliability throw item analyses [4]. Instead of the correction for guessing it would important to use systematically the item analyses, more or less complete. I agree with Robert Thorndike when he says:

In multiple-choice tests that have four or five answer choices (or options) and liberal enough time limits to permit all or most examinees to attempt every item, a score that is simply the number of correct answers is quite satisfactory, and there is little or no gain made from correcting for guessing ([4], p. 476).

5. References

[1] A. Ribeiro, A. and L. Ribeiro, *Planificação e Avaliação do Ensino-Aprendizagem*, Universidade Aberta, Lisboa, 1989.

[2] J. Valadares & M. Graça, *Avaliando para melhorar a aprendizagem*, Plátano Edições Técnicas, Lisboa, 1998.

[3] R. Ebel, *Essentials of Educational measurement*, Prentice Hall, New Jersey, 1979.

[4] R. Thorndike, *Measurement and Evaluation in Psychology and Education*, Prentice-Hall, Inc., New Jersey, 1997.

