SOME CONSIDERATIONS ON TESTS, EXAMS AND OTHER STUDENT EVALUATIONS AND ASSESSMENTS

Washington Braga
Mechanical Engineering Department
Pontifical Catholic University of Rio de Janeiro, PUC-Rio, R. Marquês de São Vicente, 225
Rio de Janeiro, RJ, Brazil
wbraga@mec.puc-rio.br

Abstract. *This paper addresses some statistical criteria to assess and evaluate tests and examinations taken by students. Recognizing that research time is considered far more important than time spent on enhancing student learning, interested faculty need to be efficient on handling student evaluation to guarantee that time is well spent. Recently, advances on the statistical area of reliability of measurements have been used to evaluate overall learning. The paper describes some of the available criteria and presents a spreadsheet developed for analysing a priori any student evaluation. Using students´ grades in ten questions and several statistical criteria, the spreadsheet analysis those grades and offers interesting conclusions on the internal consistency of individual questions and the complete examination. Results are analysed and some conclusions are drawn.*

Keywords*: engineering education, assessment, evaluation, reliability of measurements, enhancing learning*

1. Introduction

Undoubtedly, a large part of the workload of any university professor is directly related to the difficult task of assessing and evaluating student learning (Oosterhof, 2003). For a professor-educator, conscious of his/her responsibilities towards student effective learning, this turns out to be even harder. He/She believes that enhancing learning depends on many aspects such as the use of active learning techniques (e.g. Prince, 2004) and of effective feedbacks, both equally important to the development of a strong conceptual framework by the students for the course material. Quite often, this turns out to be an impossible mission, mainly on research universities that traditionally consider research as being the only sure way to promotion. Since time spent on enhancing student learning implies less time dedicated to pure academic research, faculty involvement with education is definitely considered non-productive, e.g. Cintra (1980). It seems that transmission of knowledge is to a great extent treated as a subproduct of the faculty workload. Consequently, it is not surprising that most of the faculty have almost no information on recent advances in engineering education and probably all of them so far have had no training whatsoever on educational issues. They teach the way they were taught. Since students still need to graduate, implicit in such cases is the feeling that the department chairperson hopes that by luck faculty happen to be quite knowledgeable on learning and teaching aspects. Faculty are considered able to properly specify not only the class syllabus and bibliography but also to formulate effective tests and examinations to help students evaluate their progress, avoiding, at least consciously, long and confusing questions, lack of data, new information technology resources and so on. Quite often, this is not so.

However, nowadays, tough economic restrictions are precluding hiring as easily as before and the increasing number of good graduate students looking for tenure track positions is pushing research scientists to prove them worthy of being called engineering educators. At the same time, having tuition costs in mind, students and their families (or the society at large, if one will) must be sure that a good engineering job is down the line, and are becoming generally less prone to have less competent instructors, even (or perhaps, specially) at the most prestigious universities. Therefore, it is becoming more evident to many that conducting good quality courses will be as relevant as conducting pure research, and for that, time should also be put on those academic issues, so far essentially neglected. For one reason or another, the fact is that, in a way, many university professors are now being involved in the quality of the time they spend in teaching load, questioning how this and that should be taught best, and, most important, how his/her students learn best. Fortunately it also seems that faculty are becoming aware that assessing their students´ learning is bceoming critical. As experience has indicated, student achievement is closely associated to teaching evaluation. Consequently, the issue refers once again to the kind of teaching we are doing.

Evaluating teaching is not a simple problem nor has a single approach. Student ratings, self-evaluations, prizes and other mechanisms are all different paths to finding out evidences of how well students have learned. Unquestionably, judging faculty members in terms of their research effectiveness is quite simple, as visible products (such as papers, number of graduate students, research proposals) are always available. Teaching evaluation in the past was something quite complicated but now it is known that it is possible to be done even noticing that it has a long-term effect on students, that is, the results do not appear as immediately as papers do. It has to be accomplished through a combination of factors including classroom evaluations and others. One such aspect, to be discussed in this paper, is faculty assurance that student evaluations are conveniently evaluated.

As the number of good engineering positions seems to be increasing at a slower pace than the number of graduating students, having a degree from a university well ranked at CAPES - the Brazilian governmental agency that supports

university level studies and ranks universities - is also an important asset. However, having good grades is at least as important as that, and occasionally pressure from student bodies on faculty comes into play. In many instances, faculty facing pressure from both sides (number of publications and time to prepare and grade adequate tests) decide to make easier evaluations, many times simple multiple choice examinations, using previously given and corrected homework exercises, old exam questions and similar easy fix solutions. This is quite attractive at all university campi worldwide in which academic success is affected only (or preferentially) by the intensity and the quantity (not the quality) of the developed research, nor the quality (effectiveness, if one prefers) of the faculty teaching. Having more time to research is mandatory today. Interestingly, this is not only a Brazilian issue. For instance, Mansfield (2001), from Harvard University, has given up to the pressure and started to give two grades to each student. The real one, given only to each individual, and another, public, that is sent to the Registrar and was called ironic grade by him. Students still know exactly how well they are ranked in the class but are facing no penalty whatsoever. Definitely, a more interesting (perhaps not that popular) solution is the inclusion of the whole class grade point average, class gpa, into each individual academic record or the student rank, a situation used in many countries but not in Brazil.

Although to this point in Brazil the direct influence of grades in hiring of recent graduates is not significant, experience indicates that grade competition does exist. It is common knowledge that in any university department or school there are good instructors that tend to give lower grades (sometimes, even flunk students) than others, and many times those turn out to be less popular among the students. We are all used to hear complaints from students concerning differences in course grading policies and teaching quality, depending on the instructor. Naturally, this is an oversimplification of the problem, as it considers class room methodology, workload and other factors being exactly the same, which is not exactly true. In fact, it is inevitable that students in some classes turn out to be luckier (academically speaking) than others. However, regardless of the pressure from the students, it is our function as instructors to guarantee that the quality of learning in Thermodynamics, let's say, is roughly the same, regardless of the instructor and his/her fundamental freedom of teaching and methodology.

In this paper, some criteria to help faculty ensure the same level of learning will be presented. Among many others, criteria on the minimum requirement during test elaboration and, most important, on the a posteriori analysis of the evaluation itself will be discussed. Clearly, in no way should an extensive discussion on all aspects of the academic evaluation be expected. For instance, collaborative and cooperative work, oral evaluations, projects, lab, presentations, internet education and many others will not be discussed. On the other way round, the analysis will be presented based on standard (multiple choice or not) tests. This is so not because these are considered more important, but simply because these are more suitable to statistic analysis and the known fact that, at large, faculty use them, not others.

2. Some reasons for evaluating

Once it is understood that evaluation and assessment are two cumbersome concepts, one may inquire why it shouldn't just be dropped all together. Besides the obvious responsibility towards ensuring a minimum level of understanding over the topics in the syllabus, educators such as Esteban (1999) declare that the lack of a grading system does not satisfy the students that typically enjoy being known as smart boys and girls, as nerds, following the classic stereotype that connects good grades to intelligence and success. That is, according to her, ranking is very important to them. Nonetheless, it is also often heard students saying that "examinations are not indicators of knowledge level", that during exam, they enter in panic and, therefore, failure is not indicative of anything, or "only those few topics I new nothing about were asked", and other similar phrases. To be fully honest, it is well known that there is no assurance whatsoever that a student receiving the highest grade (in Brazil, a 10 point scale is used) will definetely know twice as much as a student receiving a 5.0 grade. According to Felder (1993), it is now recognized that students having learning styles compatible to the teaching style of faculty have a much greater probability of retaining information better, applying them more efficiently and displaying a better retention rate provided comparison is done with students facing incompatibilities of this sort. Consequently, it becomes clear that some students have better chances with one methodology than others. Definitely, a series of high grades indicates an interesting pattern and for the present author this is one of the most relevant reasons for having several evaluations instead of just one. After all, it is important to find out how well students master the subjects, nothing else.

Therefore, it may be concluded that having a well balanced evaluation and assessment criterion is extremely important due to its intrinsic capability to identify weaknesses on the students´ understanding, allowing early treatment and possibly its cure. Most certainly, it is faculty's responsibility to indicate progression and any competency his/her students may have, but it is becoming fundamental to let them know if they are facing problems, the kind of their problems and paths to overcome those, Parker et all (2001). Now such feedback is considered among engineering educators to be extremely helpful to develop continuous learning habits, something that requires a good part of self-esteem and belief that it is easy to be developed at home, may be developed during university years but it is practically impossible to be developed at the work environment where tough and unfair competition is usual. Following this analysis, time spent with evaluation is well spent because all results must allow the establishment of comparative merits, even taking for granted that evaluation is an inexact art, prone to partial judgments, prejudice and some parts of subjective criteria. According to Huba & Freed (2000) and Muirhead (2002), among others, one of the most

fundamental purposes of any evaluation procedure is to give information capable of enhancing future educational experiences. In order for this to happen, it is imperative that relevant questions are formulated which demands time for preparation and correction.

3. An evaluation system

Following Felder's work (e.g. 1993), we learned about the different learning styles from both students (learning) and faculty (teaching). Therefore, to avoid biased evaluations towards students having learning styles compatible with the teaching styles of the faculty, it is now widely known that evaluation systems must consider several types of evaluation, not only plain written tests and exams. A more generalized system will consider also student presentations, challenging exercises, collaborative work, short essays and articles, and others, turning the learning process into a continuous task, not a discontinuous event that happens only close to formal exams. Students may learn or, perhaps, learn again the pleasures of discovering new information, something that was familiar at fundamental school and was lost during their academic journey. To help them keep their jobs for the coming years, they must be motivated to learn by themselves instead of having just good grades and passing in a course. As a matter of fact, succeeding is not anymore dependent on the student ability to find out the right, the correct answer as some of us, university professors, would like it to be.

One of the key factors for success is to make clear for all students which are the evaluation criteria to be adopted, since this fully alleviates tensions and anxieties. During a few terms, after showing students the evaluation criterion, the present author let them choose the weighting factors, within certain limits, for sure. In theory, they could decide if the final examination would weight more than the other exams or even the class projects or the Internet participation. This is in total agreement with the learning centered education and it started with the invitation to let them identify their own learning style, using an inquiry developed by Soloman & Felder (2005) and another one still available through the Internet, Braga (2005a). To the best of this author's knowledge, this was a very interesting experience. Unfortunately, the experience was not repeated in other courses (or at least, in the majority of them) and in a way, it held no further consequence. In any way, it was evident that knowing better how they prefer studying is something worth.

Below, some successful criteria for the development and construction of efficient tests and exams will be discussed, having in mind that they are part of a much larger evaluation criterion, as previously mentioned. One way or the other, they have been used for the last five or six years in the Mechanical Engineering Department.

4. Some routes for formulating a successful exam

Clearly, there is no sure route for succeeding but definitely there are already some recommended ones. For instance, Cohen & Wollack (2005), Svinicki (2005) and researchers of the Oklahoma University Instructional Program (2005) suggest:

- Following Voltaire, that once said that "education is what remains after one forgets everything learned," the instructor should not expect students to recall formulae, equations, etc. for more than a couple of weeks after the end of the term. Therefore, open book examinations are more than strongly suggested. Shine et al (2005) points that provided examination questions are well structured, only those candidates who fully understand the basic principles will be able to apply their knowledge;
- The instructor should identify his/her own objectives for an exam prior to selecting the very first question. For example: the evaluation will be good in any way for the next material to be covered? Knowing this or that will make any good for future courses? One should never propose a test to evaluate something from the past;
- The questions should be as simple as possible, without any superfluous or irrelevant information;
- For multiple choice questions, make all possible answers as attractive and reasonable as possible. Use mistakes and past errors to prepare new questions, exploring important topics;
- Do not include any jokes or depreciative comments, even remotely. The student's sensibility is high at such moments and an innocent comment may be disastrous;
- Try to measure or to identify what the students know, not what they do not know;
- Answers from a question should not be helpful, i.e. should not contain hints for others;
- To assess the due time for an examination, the instructor should use his/her own time multiplied for a number ranging from 3 to 6, depending on the level of the course and his/her experience with the subject;
- All student answers should be analyzed. Hints on background and fundamental issues should be observed. As often as possible, assessment should be given to them on organization, answer clarity, and other functional aspects;
- Dependent questions should never be used. If that turns out to be impossible, make sure that no error propagation will be considered. During test analysis, systematic errors should be observed as they often give hint on misconceptions;
- It is good practice to correct item per item. That is, all first questions should be corrected before moving on to the second question;

- Approval criterion should be insistently discussed with students. They should find it easily. Use the Internet to publish information like this. The rule of the game should never be changed, no matter what.

5. Evaluating the quality of an examination

The statistical concept of reliability of measurements has been used to evaluate the quality of tests and examinations, e.g. Allen et al (2004). Following Wells & Wollack (2003), considering a test reliability is important for two reasons: (a) it provides a measure of the extend to which a student's score reflects random measurement error caused by many factors, as it can be inferred, and (b) the fact that it is a precursor to test validity. In other words, it is important to find out to what extent grades reflect what they suppose to do, that may be also stated as how well do the students know the material of interest?

In order to clarify this concept, an analogy close to the one used in "Reliability and Item Analysis" (2005) will be used herein. For that matter, consider a test on Thermodynamics, for instance. It is not hard to guess that each answer will inevitably display a personal component coupled to an additional one, for instance, based on how much the subject was understood, enjoyed and if it also contains some guessing. Theory indicates that a score is equal to the true evaluation plus a measurement error. For instance, if a student knows 100% of the course material but takes 8.5 as his grade, then it may be concluded that something happened. The same thing occurs if a student knows nothing and his result is 2.0, for instance. In both situations, something strange happened but the results are clearly unreliable. The additional points may be caused by many factors: (a) motivation, concentration, lucky guess, lapses of memory (if negative), (b) the chosen questions or ambiguous items and (c) counting or computational errors. They are all considered as measurement errors. Finding out these erros is not simple but this is another subject.

A valid test (one that accurately measures the domain of interest) must be reliable (allowing inferences such as if the student knows - or not – the material being tested). If a test is unreliable, it is not necessary to discuss if it is valid. In this sense, the definition of reliability is simple: it must reflect the true score, or at least, it must not reflect a large amount of error. For instance, a question such as "water vapor may be modeled as a perfect gas" would indicate a large amount of errors because of the number of possible correct answers as there are so many thermodynamic states to consider. Therefore, any answer would bring information related to prejudice but perhaps other facts. Consequently, it may be concluded that this is a bad question (or perhaps, a not well formulated question). Following the literature, one may define an index of reliability as a ratio between the true score variability and the total observed variability:

$$\mathbf{Reliability} = \sigma^2\mathbf{(true\ score)} / \sigma^2\mathbf{(total\ observed)} \tag{1}$$

As it must be expected, each question will have a particular value for its own reliability. If the error component is truly random, then it may be expected that the different components will cancel each other. In other words, the expected value or the mean of the error component will be zero. In conclusion, it may also be said that increasing the number of questions, the true score (relative to the error score) may be better reflected. Later on this paper, the Spearman-Brown prophecy formula will be introduced for this matter and further comments on the validity of the previous sentence on a broader sense will be made. In any event, a quite simple manner to assess examination reliability is the use of the test-retest concept, that is, the repetition of the same test twice. However, for the obvious reason, it is seldom used. In the present paper, alternative actions are preferred. Before discussing them, some of the statistics used to estimate the reliability of a question will be introduced.

5.1 **Cronbach´s Alpha**

Consider an examination with several questions and students. With such information, each question and also the overall exam variances may be estimated. It may be proved that the variance of a sum of any two items is the sum of the two individual variances minus its covariance, so usually, the variance of a sum is a smaller number. The proportion of the true score variance that is taken care of by the individual items may be estimated by comparing the sum of item variances with the full test variance. That is, it may be computed as:

$$\alpha = \left(\frac{\mathbf{k}}{\mathbf{k-1}}\right) \times \left[1 - \frac{\sum \mathbf{s_i^2}}{\mathbf{s_{sum}^2}}\right] \tag{2}$$

where k is the number of questions, s_i^2 is the variances for the individual question, s_{sum}^2 is the variance for the complete test and α is the most commonly used index of reliability, namely the Cronbach´s coefficient alpha. If there is no true score but only error in the items - an odd situation of course - then the variance of the sum will be the same as the sum of variances of the individual items, and coefficient alpha will be zero (as all items will be uncorrelated, and $\sum \mathbf{s_i^2} = \mathbf{s_{sum}^2}$). If all items are perfectly reliable and measure the same thing (the true score or, in other words, which

part of the material the student really understood), then there will be a complete correlation and coefficient alpha will be equal to 1. Typically, a test is considered to be reliable if alpha is greater than 0.8. Other sources consider values greater than 0.6 to be acceptable for classroom tests. In the present work, a value of 0.7 is used.

In order for a high coefficient alpha to be achieved, it is necessary that the test variance be as high as possible, which is only obtained if the grade profile is clearly defined, indicating that students were sharply differentiated. On the other hand, very long exams are bad as it may be concluded. Note, for instance, that s_{sum}^2 increases with the number of questions and that each individual s_i^2 is a number ranging from 0 to 0.25. As the former is a quadratic term, coefficient alpha increase rate may be significant. There is a formula to handle the increase in the number of questions in a straight way. It is called the Spearman-Brown prophecy formula (Wells & Wollack, 2003) and it may be used to anticipate the reliability of a longer (or a shorter) test having in mind an existing test with a certain number of questions:

$$\alpha^{new} = \frac{m \times \alpha^{old}}{1 + (m-1) \times \alpha^{old}} \tag{3}$$

In Eq. (4), m indicates the new test length divided by the old test length. For instance, if the number of questions is doubled, it results in a percentual increase given by eq. (4):

$$\text{increase} = \frac{1 - \alpha^{old}}{1 + \alpha^{old}} \% \tag{4}$$

Provided α^{old} is a positive number, the results is clearly positive. From Eq. (2), it may be seen that whenever $\sum s_i^2 > s_{sum}^2$, the resulting coefficient alpha will be negative. Nichols (2005) analyzed this and concluded that this is a sharp indication of strong measurement errors, perhaps some typing or systematic errors of some sort, as it implies a negative covariance among items (or questions). Observe that a marginal increase in the number of questions does not affect much that coefficient. On the other end, if the number of equations is significantly increased, say by a factor of 10, the net effect is confusing as $\alpha^{new} \to 1$, regardless of the questions being good or bad, a situation that does not mean much. Consequently, one may conclude that very long tests are definitely not reliable, even assuming that the quality of new the questions is similar to the quality of the previous ones.

5.2 **Split-Half Reliability**

Another way to calculate the reliability of an exam with several questions is based on splitting the test in two parts, which may or may not be randomly chosen. A simple and quite used way of doing this is selecting the group of even questions and the group of odd questions. If the evaluation is perfectly reliable, then both parts will be completely correlated. In this case, the correlation coefficient will be equal to 1.0. A less than reliable evaluation will have smaller correlation coefficients. Using the Spearman-Brown split half coefficient:

$$r_{SB} = 2r_{xy} / (1 + r_{xy}) \tag{5}$$

In the above equation, r_{SB} is the split-half reliability coefficient and r_{xy} indicates the correlation between the two halves of the scale. It may be observed that this criterion is another way of applying the well known test-retest methodology. In the present situation, both tests are done at the same time, in a quicker and cheaper way, as both halves analyze information collected at the same time.

5.3 **Exam reliability**

In theory, reliable tests are those that give repeatable results, even if the student takes them on different days. In the real world, this is not to be expected due to the presence of measurement errors, no matter how small they are. Allen (2004) cites that there are several ways of assuring test reliability: test-retest, two similar tests on the same subject, and internal consistency. This last one, implemented herein, is based on inter-item correlations and it is measured using Cronbach´s alpha.

5.4 **Single question reliability**

The concept of internal consistency may be also used for each question individually. This is done through the alpha-if-deleted criterion. That is, the same coefficient alpha is calculated considering the deletion of a particular question. A question is considered as being consistent provided the resulting coefficient alpha is less than the overall coefficient alpha for the full test, as its presence increases the overall coefficient and, therefore, the test reliability. In practice, questions are considered good provided the absolute difference between their coefficients alpha-if-deleted and the overall coefficient are smaller than 0.03.

5.5 **Discriminatory power**

Following Allen et al (2004), discrimination refers to the ability of a test to produce a wide range of scores or grades. One of the main objectives of any evaluation is to identify students according to the level of knowledge in the course subject. Therefore, any evaluation in which all (or a large portion) of the students have roughly the same scores, no matter if equal to 0, 5 or even 10 points, can not fit such definition. A simple but instructive example comes from a normal distribution. At both tails, as the number of students is small, it is easier to differentiate students in terms of their shown knowledge. However, the middle scores are much harder because they are clustered. This power may be measured by at least two ways (both implemented in the spreadsheet). The first one is through the Ferguson´s coefficient delta, a number ranging from 0 (all grades are equal) to 1 (each student has its own score). Following such criteria, a test having $\delta > 0.90$ is considered as being discriminating. In any event, it is calculated using Eq. (6):

$$\delta = \frac{(k+1)(N^2 - \Sigma f_i^{\,2})}{kN^2} \tag{6}$$

where k is the number of items, N is the number of students and f_i is each individual frequency (if 5 students got 8 and 10 got 10, then their frequencies are 5 and 10). While Ferguson´s coefficient delta relates to the overall test, it may also be important to find out if each question is discriminating and for this suffices a comparison between the best and the worst scores. That is, consider that 75% of the students with the best grades and 35% of those with the worst grades succeeded in a certain question. In such case, a discriminating index of 0.75 – 0.35 = 0.40 may be assigned to that question. The optimum cut point, according to the literature, is 27% but in practice, one uses 25%, Allen [2005]. A question is considered poorly discriminating provided its index is less than 0.20. If its index is larger than 0.40, it is considered as being discriminating.

6. **The overall effect**

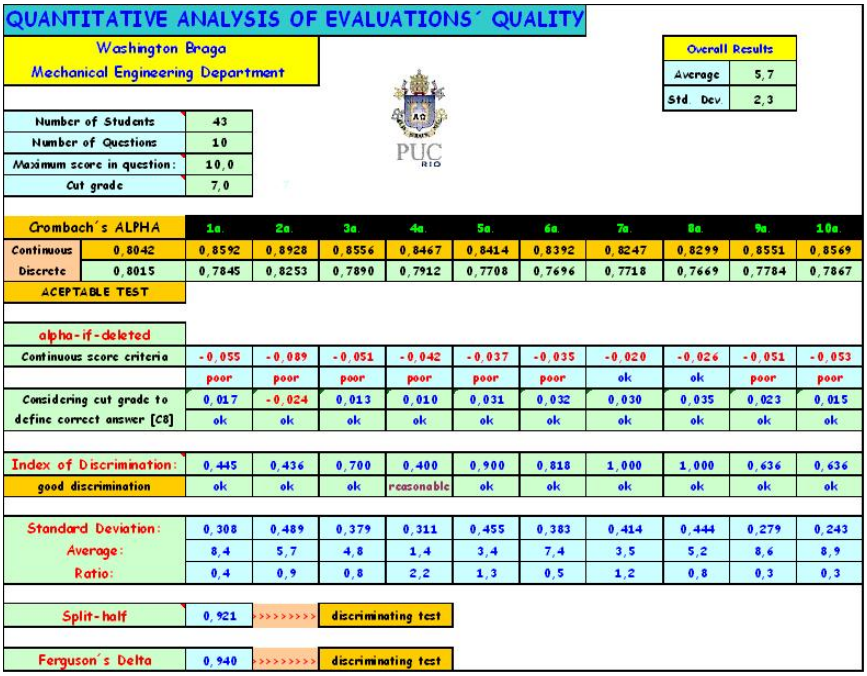
Once it was understood the existence of different criteria to evaluate tests, it was decided to group them into a spreadsheet in order to have a global analysis. The resulting spreadsheet report is shown in Figure 1. Some comments are due at this time: first of all, there are two sets of information concerning Cronbach´s coefficient alpha. The first one is obtained using scores considered as a continuous distribution from zero to ten. The second one handles scores in a more discrete way. Internally, answers are considered as correct if the assigned grade is equal or greater than the chosen grade of cut. Consequently, there are two sets of coefficient alpha-if-deleted. In the present author's experience with evaluations using such criteria, the discrete analysis is usually the fairest one, due to the many uncertainties involved in the analysis shown.

The next set of data refers to the index of discrimination. In the test shown, all but one question was truly discriminating and the other was considered reasonably discriminating (index is less than 0.7). According to this criterion, the test was quite reliable (as 9 out of 10 questions were found to be discriminating). The next information shown is the average grading (each question has a maximum score indicated by the user in the cell "maximum score in question"); the average grade and the normalized ratio between them. Following this analysis, having a large standard deviation is good sign for a good question (and consequently for a test). Therefore, increasing this ratio is interesting, at least considering the same average. Finally, two other sets of data are shown. The first indicates the result for the split-half analysis and the second indicates the Ferguson´s delta. Following both results, the test was discriminating.

Before concluding, it is instructive to find out what constitutes well or poorly proposed questions based on this analysis. According to the theory, a question is considered as being poor provided its internal consistency is low, as indicated by an increase in the overall coefficient alpha. This implies that there is no coherence between that question and the remaining of the test. The experience indicates that this may happen whenever guessing helps finding an answer or there are hints from other questions, for instance. In some situations, the discriminating index previously mentioned results negative. Since this indicates that students having the lower grades answered correctly a certain question but the students with the best grades did not, which is a unreasonable situation, questions like should be considered unfit and eliminated from the test or at least they should be avoided.

To conclude the present analysis, three examples of such questions will now be presented, both taken from a mid-term test for an undergraduate Thermodynamics course.

Q1: Consider a rigid container in which an axis is able to withdraw 550 kJ/s. What is the total work?



QUANTITATIVE ANALYSIS OF EVALUATIONS´ QUALITY

Washington Braga
Mechanical Engineering Department

Overall Results	
Average	5,7
Std Dev	2,3

Number of Students	43
Number of Questions	10
Maximum score in question:	10,0
Cut grade	7,0

Crombach´s ALPHA		1a.	2a.	3a.	4a.	5a.	6a.	7a.	8a.	9a.	10a.
Continuous	0,8042	0,8592	0,8928	0,8556	0,8467	0,8414	0,8392	0,8247	0,8299	0,8551	0,8569
Discrete	0,8015	0,7845	0,8253	0,7890	0,7912	0,7708	0,7696	0,7718	0,7669	0,7784	0,7867
ACEPTABLE TEST											
alpha-if-deleted											
Continuous score criteria		-0,055	-0,089	-0,051	-0,042	-0,037	-0,035	-0,020	-0,026	-0,051	-0,053
		poor	poor	poor	poor	poor	poor	ok	ok	poor	poor
Considering cut grade to define correct answer [C8]		0,017	-0,024	0,013	0,010	0,031	0,032	0,030	0,035	0,023	0,015
		ok	ok	ok	ok	ok	ok	ok	ok	ok	ok
Index of Discrimination:		0,445	0,436	0,700	0,400	0,900	0,818	1,000	1,000	0,636	0,636
good discrimination		ok	ok	ok	reasonable	ok	ok	ok	ok	ok	ok
Standard Deviation:		0,308	0,489	0,379	0,311	0,455	0,383	0,414	0,444	0,279	0,243
Average:		8,4	5,7	4,8	1,4	3,4	7,4	3,5	5,2	8,6	8,9
Ratio:		0,4	0,9	0,8	2,2	1,3	0,5	1,2	0,8	0,3	0,3
Split-half		0,921	>>>>>>>>	discriminating test							
Ferguson´s Delta		0,940	>>>>>>>>	discriminating test							

Fig. 1: Typical display from the spreadsheet used

Comments: students were supposed to know that in a rigid container, the volume is constant and no moving boundary work is possible. Therefore, the only type of work is associated to the axis. Since energy is going out, the work is considered to be positive, according to Thermodynamics convention. That is, the only possible answers were the correct one or the wrong one. In this case, the simplicity of the question resulted on a poorly discriminating answer as indicated by a coefficient alpha-if-deleted greater (0.725) than the overall coefficient alpha (0.702). Only 33% of the best students succeeded and none of the poorly ranked students did so.
Q2: A rigid vessel contains a known volume of NH_3 at a certain pressure and temperature. Estimate its mass.
Comments: The students had to obtain the specific volume of the substance. Prior to that, they had to find out if the substance could be modeled by the perfect gas equation. In that situation, it could not. However, many students observed the existence of vapor tables for this particular substance and shortcut this previous analysis. In the end, the coefficient alpha-if-deleted was slight higher than the overall coefficient alpha, the discriminating index was just 33% (83.3% of the best students answered correctly and 50% of the poorly ranked ones).
Q2: Given two thermodynamic states and the process between them, what is the total heat transfer?
Comments: students should use water vapor tables to find out the internal energy variation between those two states. They should be able to calculate the work transferred between those states and finally apply the 1st Law of Thermodynamics to calculate the total heat transfer, indicating further the direction of the transfer. The overall coefficient alpha was 0.702 and the coefficient alpha-if-deleted for this question was 0.593. It happened that 100% of the best students were able to answer correctly and none of the poorly ranked students were not. So it seems that having involving questions in which there is no space for shortcuts or plain guessing is important.

7. **Final Remarks**

Clearly, there is a greater demand, these days, for engineering faculty to enhance student learning through the use of many interesting methodologies based on active learning. However, to handle properly the ever increasing research time, faculty must be sure that they are using their time efficiently. Since up to this point student learning is best viewed through any grading system, it is becoming important that a good measurement rule exists for examinations and tests. The present paper has shown several criteria, already being used, which give faculty a good grasp on any evaluation method. No such criterion is fool proof and there is certainly much more to be analyzed along the statistical field of reliability but, so far, the results are interesting. Those criteria were implemented on a spreadsheet, freely available through the Internet at the author's homepage, Braga (2005b). Although the results from such spreadsheet are quite obvious, there is certainly a much more relevant lesson to be learned by each instructor interested in helping his/her students to become better engineers: how to know, beforehand, if an examination is reasonable, considering that many times it will approve or not the students. Clearly, the experience is crucial in such cases and a handful of information learned from experience was already discussed. However, the mere existence of these (and perhaps, many other) criteria and its easiness of using a spreadsheet has definitely helped the present author to prepare more reasonable evaluations. At least, it may be said that he is being much more careful during exam preparation.

8. Acknowledgements

The author wishes to thank PUC-Rio for supporting the research herein presented and to Prof. P. M. Gouvêa, Metrology Department of PUC-Rio, who read carefully the manuscript and made invaluable contributions, making this paper more reliable.

9. References

Allen, K., Stone A., Rhoads T.R. & Murphy T.J., "The Statistics Concepts Inventory: Developing a Valid and Reliable Instrument", Proceedings of the Annual Conference and Exposition of the American Society for Engineering Education, ASEE, 2004.

Allen, K.,"Explaining Cronbach´s Alpha", available at http://coecs.ou.edu/sci under the item "publications", accessed on May 2, 2005.

Braga, W., "Pré-requisitos para MEC 1340: Transmissão de Calor, available at http://wwwusers.rdc.puc-rio.br/wbraga/primer.htm, accessed on May 2, 2005, in Portuguese.

Braga, W., Internet home page, available at http://wwwusers.rdc.puc-rio.br/wbraga/hpn.htm, accessed on May2, 2005, in Portuguese.

Centra, J. A., "Determining Faculty Effectiveness", Jossey-Bass Publishers, San Francisco, California, 1980.

Cohen, A.S. & Wollack, J.A., "Helpful Tips for Creating Reliable and Valid Classroom Testes", Testing & Evaluation Services, University of Wisconsin-Madison, available at http://wiscinfo.doit.wisc.edu/exams/instructional_support.htm, accessed on May 2, 2005.

Esteban, M.T., "A avaliação no cotidiano escolar", em "A avaliação: uma prática em busca de novos sentidos", Maria Teresa Esteban (org). DP&A Editora, Rio de Janeiro, 1999, in Portuguese.

Felder R.M., "Reaching the Second Tier: learning and teaching styles in College Science Education", J. College Science Teaching, vol 23, no. 5, pp. 286-290, 1993.

Huba, M.E. & Freed, J. E., "Learner-centered assessment on college campuses: Shifting the focus from teaching to learning", Allyn & Bacon Editors, Boston, Ma, 2000.

Kelley, T., "The Selection of Upper and Lower Groups for the Validation of Test Itens", Journal of Educational Psychology, vol 30, pp 17-24, 1939.

Kuder & Richardson, "The Theory of the Estimation of Test Reliability", Psychometrika, vol 2, no. 3, 1937

Mansfield, H.C., "All shall have prizes", The Economist, April 12, 2001.

Muirhead, B., "Relevant Assessment Strategies for Online Colleges & Universities", available at http://www.usdla.org/html/journal/FEB02_Issue/article04.html, February 2002, Vol. 16, no. 1, accessed on May 2, 2005.

Nichols, D.P., "My Coeffient α is Negative!", available at http://www.ats.ucla.edu/stat/spss/library/negalpha.htm, accessed on May 2, 2005.

Oosterhof, A., "Developing and Using Classroom Assessments", 3rd Edition, Merrill Prentice Hall, 2003.

Parker, P., Fleming, P., Beyerlein S., Apple D. & Krumsieg K., "Differentiating Assessment from Evaluation as Continuous Improvement Tools", 31st ASEE/IEEE Frontiers in Education Conference, October 2001, Reno, NV, United States.

Prince, M., "Does Active Learning Work? A review of the research", Journal of Engineering Education, ASEE, Vol. 93, no. 3, pp 223-231, July 2004.

"Reliability and Item Analysis", available at http://www.statsoftinc.com/textbook/streliab.html#cronbach, 2003, accessed on May 2, 2005.

Shine, S., Kiravu C. & Astley, J., "In defence of open-book engineering degree examinations", Int. J. of Mechanical Engineering Education, volume 32, issue 3, pps 197-211, July 2004.

Soloman, B. & Felder, R., "Index of Learning Styles Questionnaire", available at http://www.engr.ncsu.edu/learningstyles/ilsweb.html, accessed on May 2, 2005.

Svinicki, M., "Test Construction: Some Practical Ideas", available at http://www.utexas.edu/academic/cte/sourcebook/tests.pdf, accessed on May 2, 2005.

"Test Construction and Assessment", in "Ideas on Teaching", Instructional Development Program, University of Oklahoma, available at http://www.ou.edu/idp/tips/ideas/quick10.html, accessed on May 2, 2005.

Wells, C.S. & Wollack, J.A., "An Instructor´s Guide to Understanding Test Reliability", available at http://wiscinfo.doit.wisc.edu/exams/instructional_support.htm, November 2003, accessed on May 2, 2005.

10. Responsibility notice

The author is the only responsible for the printed material included in this paper.