Spectral Quality of University Standardized Tests

Development of edumetrical indices for the analysis of the spectral quality of higher education standardized tests and application to the MOHICAN check up '99 tests

Jean-Luc GILLES

State University of Liège - Belgium

Ph.D. dissertation abstract – January 2002

For several decades almost all European university institutions have been faced with a strong increase in their number of students whereas the allocated budgets have not increased in proportion (Gibbs & Jenkins, 1992). The universities of the French Community of Belgium do not escape this heavy tendency: compared with the 1972 situation, the number of students has increased to 150% and, in constant francs, subsidies have remained unchanged (Debry & al., 1998). This situation leads in the first cycles of studies, where the students are most numerous, to a massive use of standardized examinations with multiple choice questions (MCQ) which makes the assessment of great groups of students possible within reasonable time delays.

The confidence degrees technique associated to the MCQ makes it possible to bypass the "binary" character of students assessment performance (the selected proposal is either correct, or incorrect) provided that a series of methodological rules are followed called "*admissible probability measurement procedures*" by Shufford & al. (1966). Among those rules: *ask the students to express their confidence in a numerical (probabilistic) scale*. The one adopted offers 6 degrees of subtlety: 0%, 20%, 40%, 60%, 80% and 100% of confidence. By inviting the student to accompany his choice by the percentage of chances to be correct he attributes to his answer, we allow more subtleties in the analysis of his performances. At one extreme, the choice of a wrong answer accompanied by the maximal percentage of certainty (100%) presents the worst situation, in which the student provides an erroneous answer by estimating that he has a maximal chances to be correct. On the other hand, the student who answers correctly with a maximal certainty shows an assured performance. Between these two extremes, other zones can be identified in the "spectral analysis" of the performances. Jans & Leclercq (1999) propose a specific terminology. They distinguish "*ignorance*" (correct answer and weak certainty), from "*partial knowledge*" (correct answer and average certainty), and "*perfect knowledge*" (correct answer and high certainty). Such spectral gradations were also considered by these authors in the case of an incorrect answer ("*mistaken knowledge*" and " *fallacious knowledge*" or "*dangerous knowledge*")

Usually, the confidence percentages which accompany the MCQ answers are used to deliver more subtler feedback on each student's spectral performances. The innovative aspect of our approach lies in the fact that we have exploited the confidence percentages provided by the students to provide spectral information on the quality of the questions (as opposed to information on the quality of students performances). Our research thus led to the development of a series of original indices for the analysis of assessments' spectral quality. These spectral indices are intended to be used when the assessor must highlight problematic MCQ and, within those, the proposals which contain anomalies.

Our starting intuition for the construction of these new indices is as follows: logically the students who answer a question correctly should provide percentages of certainty higher than the students who answer incorrectly. Thus, for a multiple choice question which functions normally from the point of view of the certainty percentages use, we should observe among the students who choose the correct answer a tendency to answer with higher percentages of certainty and, in parallel, among the students who choose a wrong proposal a tendency to answer with lower percentages of certainty. We will then say that there is "*spectral coherence*". In the case where this situation does not arise, for example when students choose higher certainty percentages for one of the incorrect answers rather than for the correct answer, we are confronted with a problem of inconsistency in the use of the percentages of certainty, we will then speak of "*spectral inconsistency*".

To measure spectral coherence we created two new types of indices starting from the classical point bi serial correlation coefficient (*classical rpbis*) calculation principle. Let us recall that in the case of the *rpbis*, the choices or the rejections (1 or 0) of each proposal of a MCQ are correlated with the numbers of correct answers obtained to the whole test. The *classical rpbis* makes it possible to evaluate up to what extent each proposed alternatives solutions of each MCQ discriminates the students according to the criterion of the number of correct answers. Logically, one expects

that the students who collect a high number of correct answers tend to choose the correct answer for a given question and that the students which collect a lower number have a tendency to choose an incorrect proposal.

The two new types of spectral coherence indices of measurement are: (1) the *Spectral Contrasted rpbis* (*rpbis SC*) and (2) the *Spectral Contrasted rpbis* calculated after *Turbo analysis* (*rpbis SCT*). During a former research we had already used information related to the confidence degrees to calculate a new type of *rpbis*, the *Spectral rpbis* or *rpbis S* (Gilles, 1998). The *rpbis S* was developed in order to analyze the tendency to use higher certainty in case of correct answers than in case of incorrect answers. Within the framework of this thesis we propose a first improvement of the *rpbis S* by implementing a "*contrasted treatment*" for the MCQ incorrect proposals.

We use the name "*rpbis SC*" to indicate the *rpbis S* is constructed with a "*Contrasted treatment*" which consists in the computation of the *rpbis SC* of an incorrect answer by using the data of the students who choose this wrong response in contrast with the data restricted to the students who choose the correct answer excluding the data from the students who choose an other incorrect answer. The advantage lies in the elimination of the data of the students who have chosen the other incorrect answers. This avoids introducing into the measurement of spectral coherence, the "background noise" generated by the data of the other incorrect answers.

The principle of "*turbo analysis*", consists in operating a selection in the data used for *rpbis SC* calculation on the basis of the level of realism reached by the students. We can thus increase the reliability of information related to the spectral indices by restricting the data to the students who make less errors in their self-assessments (in their use of certainty percentages). We measure the level of errors of self-assessments made by the students by using the index of realism which varies from 0 to 100 (Leclercq & al, 2000). The name *rpbis SCT* indicates *rpbis SC* calculated with a *T*urbo analysis. The word "turbo" refers to the rise to power of the instrument in terms of the quality of information provided as one restricts oneself progressively to data from sets of students who make less and less errors in their self-assessments. We mention in the index's name the threshold of realism used to select the data. For instance *rpbis SCT80* was calculated starting from the data of the students whose realism is equal to or higher than 80 (those who make between 0% and 20% of errors in their self-assessments).

The new indices *rpbis SC* and *rpbis SCT*, which are in the heart of this research, are designed for the detection of problems located at the "*alternatives level*" inside the MCQ. We also adapted other spectral indices initially planned for the analysis of students performances so that these indices deliver information on the MCQ performances, at a "*question level*". On one hand, the index of *Realization of the predictions by question* (*Rq*) or the quantity of errors of self-assessment contained in the results for a question and, on the other hand, the index of *Centration per question* (*Cq*) or the level of over or under confidence in the results for a question. Lastly, we also measured the *level of spectral coherence of a question* (*NCSq*) by taking into account the *rpbis SC* of the various proposals of a MCQ. The "*turbo analysis*" principle was also applied to the calculation of these spectral indices.

We tried to test these new spectral indices designed for detection of suspect alternatives within suspect MCQ using the data of several thousands of answers and certainty percentages collected during the ten standardized tests of the MOHICAN project (Leclercq & al., 2001). They consisted in ten tests of knowledge of the principal subjects at the level of the end of secondary education which were submitted to groups of students entering first year studies in eight of the nine university institutions of the French Community of Belgium (the number of questioned students varied between 1.392 and 3.846 according to tests). These standardized tests consisted of MCQ for which students were invited to choose one answer and systematically accompany it with a percentage of certainty. The students had, not only to state which was their correct proposal, but also which was the percentage of chances to be correct that they granted to each one of their answers. No academic scores were delivered for these (Check up) MOHICAN tests (each student received a diagnostic feedback and the assessors a global feedback). Anonymity was guaranteed. The choice of the percentages of certainty was thus influenced neither by a scale of tariff of points nor by granting a final score which could have affected the later academic course of the student.

The total number of MCQ for the ten MOHICAN tests was 173. For two MCQ among them, the 3rd and the 20th question of the general knowledge test in History and Socio-Economy, the values obtained with the *spectral rpbis* indicate situations of marked spectral inconsistency, the students tending to give lower percentages of certainty for the answers considered as correct and higher percentages of certainty for the incorrect answers. Studying the proposals of the two problematic MCQ by using the *classical rpbis* indices, we notice that the two MCQ do not function correctly from the point of view of classical discrimination. When we ask for the opinion of the experts of the contents, they confirm that the two MCQ display problematic results: for one of the questions a wrong proposal could also be regarded as being a correct answer and for the other, there is an error in the encoding of the correct answer. For these two questions there is thus convergence of three different perspectives: (1) that of the *classical rpbis*, (2) that of the experts and, (3) that of the spectral coherence measured using the *spectral rpbis*. In the case of the MOHICAN tests, the

spectral analysis thus allows to highlight two questions that a more qualitative analysis (posterior opinions of the experts) as well as a classical analysis of discrimination (*classical rpbis*) also indicate as questions with problems.

Does the spectral analysis make it possible to perform better than the classical analysis of discrimination (*classical rpbis*) for locating the problematic questions and the proposals which contain anomalies? This issue has been tested. We analyzed the 173 MCQ of the 10 MOHICAN tests by using the *spectral rpbis* (*rpbis SC*, *rpbis SCT80* and *rpbis SCT90*) as well as the *classical rpbis* indices. We also reviewed the comments formulated by the experts of the contents in connection with each question. From these analyses it comes out that in addition to the two MCQ already discovered previously, 14 other questions are highlighted. The *classical rpbis* indicates anomalies in each one of these 14 MCQ. Six MCQ present abnormal values at *rpbis SC*. Only one MCQ obtains an abnormal *rpbis SCT80*. No MCQ obtains an abnormal *rpbis SCT90*. Among these 14 questions, only 3 are pointed as problematic by the experts.

With regard to the three questions singled out by the experts, they lead to a set of proposals that only the *rpbis SC* designates, whereas for the *rpbis SCT80* and the *rpbis SCT90*, they are not highlighted. As far as the *classical rpbis* is concerned one only of the two problematic proposals for only one of the three questions appears. Therefore, from the point of view of "detection", *rpbis SC* were more effective to highlight the problems raised by the experts than were the other indices.

This analysis shows that the various types of rpbis induce also "false alarms", abnormal values collected by proposals whereas the experts of the contents do not detect particular problems. From this point of view, the *rpbis SC*, with 7 false alarms, is less effective than the *rpbis SCT80* which causes only one of them and less than the *rpbis SCT90* which starts none, but *rpbis SCT80* and *rpbis SCT90* fail by "*undetecting*" the three questions pointed by the experts. On the other hand the *rpbis SC* starts less false alarms than the *classical rpbis* which has 10 false alarms. These qualities of lower "*undetection*" and less "false alarms" are crucial when the question arises of highlighting a problematic MCQ.

When we correct the anomalies contained in certain answers within the MCQ, we can, from now on, not only evaluate the spectral impact of these corrections on the answers ("*alternatives level*"), but also on the whole question by comparing the values obtained with the indices of *NCSq*, *Rq* and *Cq* before and after changes are operated (at the "*question level*"). We did this for the two most problematic questions of the general knowledge test in History and Socio Economy and quantified the gains in spectral coherence. The improvement of the spectral coherence of the test was also measured by calculating the average values of the spectral indices at the "*question level*". These average indices thus made it possible to evaluate the spectral impact on a third "*test level*" of the assessment. In parallel, we also observed an improvement of the fidelity of the test using the classical Cronbach's alpha coefficient.

Using the spectral indices developed within the framework of our thesis and usable with three levels of spectral analysis: "ALTERNATIVES", "MCQ" and "TEST", we open a new way for the analysis of the quality of standardized tests and their regulation. We are now able to evaluate the spectral quality of higher education standardized tests using confidence degrees, to highlight possible anomalies in the questions; and, after corrections, to evaluate the spectral impact of the improvements. This is the main contribution of our thesis to the improvement of procedures that control the quality of standardized tests and, by extension, to the improvement of the reliability of the students scores, which, *in fine*, constitutes the stake of our edumetric concerns.

* * *