SAGE Open
http://sgo.sagepub.com/

Calibrating Questionnaires by Psychometric Analysis to Evaluate Knowledge

J. Manuel Gómez-Soberón, M. Consolación Gómez-Soberón, Ramón Corral-Higuera, S. Paola Arredondo-Rea, J. Luis Almaral-Sánchez and F. Guadalupe Cabrera-Covarrubias

SAGE Open 2013 3:
DOI: 10.1177/2158244013499159

Published by:
SAGE
http://www.sagepublications.com

SAGE Open
July-September 2013: 1–14
© The Author(s) 2013
DOI: 10.1177/2158244013499159
sgo.sagepub.com

Calibrating Questionnaires by Psychometric Analysis to Evaluate Knowledge

J. Manuel Gómez-Soberón[1], M. Consolación Gómez-Soberón[2], Ramón Corral-Higuera[3], S. Paola Arredondo-Rea[3], J. Luis Almaral-Sánchez[3], and F. Guadalupe Cabrera-Covarrubias[1]

Abstract

The experience achieved using the tool "Questionnaires," available inside the Virtual Campus of an architectural engineering school in northeast Spain, is presented. "Questionnaires" is a mechanism/tool adequate and simple to evaluate the knowledge level achieved by students. This work shows and identifies the control indices of adaptation for questionnaires, such as the Facility Index, the standard deviation, the Discrimination Index, and the Discrimination Coefficient. From these parameters, educational performance is inferred, identified, and predicted. The conclusions of this work will allow the modification of deficient knowledge-evaluation practices and the identification of needs for specific groups or for students with particular requirements, thus making it feasible to apply these parameters with a guarantee of success in similar evaluation processes.

Keywords

calibration of a questionnaire, evaluation of the questions, questionnaires in teaching

Introduction

European universities nowadays are prompting conceptual and structural changes inside the European educational space. These universities should be about the work of professors, the form by which the knowledge should be transmitted, easier ways of learning for students, and finally, achieving satisfaction according to the social context of a competent education as required by society and the educational institution (Goldston, Dantzler, Day, & Webb, 2012; Veiga & Amaral, 2009). To achieve these objectives, higher education institutions in Spain in recent years have started improvement processes in their teaching practices (J. M. Gómez-Soberón, 2009; J. M. Gómez-Soberón & Gómez-Soberón, 2007; J. M. Gómez-Soberón, Gómez-Soberón, & Gómez-Soberón, 2009).

Within the outlined context, in which the principal challenge for educational development is the generation of mechanisms or evaluation systems that produce relevant information on what is taught and learned in a way that is effective in schools, we have begun to deliberate on all our educational processes. For this, we have incorporated the statistical use of questionnaires that allow us to define parametric indices about the learning of students. These questionnaires may be seen as a new educational tool that responds to the current demand for the analysis of learning and the redirection of possible tendencies or undesirable deviations in students.

There are published studies concerning the application of evaluation methods to assess possible improvements that foster learning in students (Bowles, 2008; Britt, McCall, Austin, & Piterman, 2007; Fančovičová & Prokop, 2010; Jaeger, 1998; Mowder & Shamah, 2011; Murayama, Zhou, & Nesbit, 2009; Rivero, Martínez-Pampliega, & Olson, 2010; Zhang, 2011). This research presents evaluation questionnaires as tools by which data processing systems can be an effective and convenient strategy to reinforce student learning. Some of the advantages of these systems are the efficient management of results, the speed by which the evaluation can be performed, and avoidance of the use of paper questionnaires (Shepard, 2006). However, there are some objections to the implementation of such systems. These objections concern the confidentiality of the identity of the student, the subsequent use of the information and its possible impact on the educational process (Burden, 2008; Nulty, 2008), and the implications of the use of such systems as the

[1] Universidad Politécnica de Cataluña, Barcelona, Spain
[2] Universidad Autónoma Metropolitana, México D. F., Mexico
[3] Universidad Autónoma de Sinaloa, Mexico

Corresponding Author:
J. Manuel Gómez-Soberón, Escuela Politécnica Superior de Edificación de Barcelona, Universidad Politécnica de Cataluña, Av. Doctor Marañón 44-50, CP 08028, Barcelona, Spain.
Email: josemanuel.gomez@upc.edu

sole criterion for information in the learning process of students (Garfield, 2003).

There are few bibliographical references regarding the analysis of information generated from test-type questionnaire evaluations. The analysis of such information gives professors an idea of the meaning of the results and how such results could be useful for professors and students (Blanco & Ginovart, 2009, 2010; Blanco, Estela, Ginovart, & Saà, 2009; J. M. Gómez-Soberón, Bosch-González, Marín-i-Gordi, Gómez-Soberón, & Gómez-Soberón, 2010; Romero, Ventura, & Salcines, 2008).

For the design of evaluation tests, the writing of multiple-option questions is a specialized task that requires personnel with experience and training. If such questions are adequately elaborated, they will be able to measure complex educational abilities (depending on the knowledge and experience of the person writing them; Esquivel, 2000). The description of items or questions implies verifying the relation between the item and the content supposedly being measured by the item. This verification is considered a central part of test validation processes (it is usual to carry out such verification by student-professor feedback). Therefore, to confirm the validity of the test, the information about the content and the clarity and comprehension of the items are important.

In light of the previous statements, it is necessary to perform statistical analyses to determine the characterization indices of items (difficulty, correlation item-total score, discrimination, and answer frequency according to option) and to select the adequate items for evaluation by theory of tests with reference to norms, establishing as ideal questions those close to 50% of difficulty and a discrimination above 0.40, and a correlation item-total score that is positive and significantly higher than zero (Esquivel, 2000).

However, both processes, namely, the analysis of test results and the application of questionnaires, require extra effort from teachers, which causes lost time from teaching or assessment. The main causes of lack of interest would be the following:

1. Resistance to new evaluation tools by the traditional professor,
2. Refusal to investigate in situations that are not expected to be repeated,
3. "Extra" time in educational tasks,
4. Possibility of generating additional resources, which cannot be assumed.

It is important to note, however, that there currently are tools and calculation processes that make possible the analysis of multiple processes, the generation of simulations, or the validation of prediction hypothesis about guidelines inside the education field, all of which can become very useful in specific cases or individuals, if they consider theoretical and mathematical principles on which they are based and applied (Hutchison, 2009).

Table 1. Students in the Course of Study and Grades Obtained.

Year	Course	No. of students	HM	E	N	A	U-A	A-C
2002/2003	1Q	179	0	0	1	80	91	7
	2Q	209	0	0	1	77	120	11
2003/2004	1Q	339	0	0	2	88	212	37
	2Q	357	0	0	2	114	210	31
2004/2005	1Q	367	0	0	1	85	250	31
	2Q	374	0	1	31	123	211	8

Note. HM = honorable mention; E = excellent; N = notable; A = approved; U-A = unapproved; A-C = abandoned curse.

Educational Framework and Study Participants

"Constructions of concrete" is the course studied in the present work; this course is part of the studies toward a university technical degree. This is a 4-month course in the 2nd year of study (Obligatory in the Curriculum Block). It takes place in the 4-month term 2B, and it consists of six credits (not European Credits Transfer System credits), subdivided into 4.5 theoretical credits and 1.5 practice credits.

The subject is simultaneously given to four groups in all the 4-month terms (1Q: autumn; 2Q: spring): two groups of students in the morning (Groups 1M and 2M) and two groups of students in the afternoon (Groups 3T and 4T).

Table 1 presents the number of students who have taken this course in recent years. It can be observed that the number of students registered has been increasing with the time, resulting in educational problems, such as extra work by professors, decrease in teaching quality, difficulty in evaluation using traditional systems, and so on.

It can be said that in this discipline, a high number of students fail and enroll in the course multiple times. Study participants are all in their first attempt at this course (Escuela Politécnica Superior de Edificación de Barcelona, 2009).

Research Design

To design the evaluation analysis system and to obtain the control indices to utilize in this work, some general criteria and practical recommendations were followed to guarantee a correct application of the work and to avoid bias with an incorrect use of the work (Myford & Engelhard, 2001; Ravela, 2000; Tiana, 1997). In this way, the information on the person's process, the statistics of the answers that were used, and the analysis and the percentages of the evaluation questionnaires (multiple for each student) were accomplished in the tool "Questionnaires" of the Virtual Campus (Dougiamas).

As the starting point in the process of analyzing the concerned data, the border of sampling was delimited to the course and group submitted for analysis, considering the following aspects:

Table 2. Evaluation Procedure of the Subject.

Thematic content (module)	Contents (%)	Technical of evaluation	Final score (%)
1, 2, 3, 4, 5 6, 7, and 8	50	Practical Work No. 1	7.50
		1st partial examination	40
9, 10, 11, and 12	50	Practical Work No. 2	7.50
9	15	Test 1[a]	5
10	15	Test 2[a]	5
9, 10, 11, and 12	50	2nd partial examination	45

[a] Optional, not score in the final grade.

Table 3. Bloom's Taxonomy of Evaluative Test.

Level of knowledge	Subcategory	No. of questions, Test 1	No. of questions, Test 2
1	Knowledge	1	4
2	Understanding	9	13
3	Application	7	5
4	Analysis	9	9
5	Synthesis	1	0
6	Evaluation	3	10
	Total	30	41

1. Analysis period: 2008/09
2. Course: 1Q
3. Group where the analysis is done: 4T

The motives that are reduced by the analysis of this work to the previous variable (period, course, and group) are the only time available for classes and teachers who agreed to participate, and classes with approval by the school: experience initial calibration, verification of their suitability, and so on, considering that this is a pilot. The study group represents the total number of students in the group and course, and therefore constitutes a stratified response to a type of nonparametric sampling.

The analysis presented pertains to data processed and extracted from specific evaluations (two midterm exams), from two works done by students, and from two tests (multiple-option and paired-test type; Tuparov & Dureva-Tuparova, 2008). Table 2 shows the evaluations of previous techniques.

The specific evaluations were individual, consisting of solving graphic-conceptual problems. The activities developed by the students involved the resolution of real cases, with applications related to topics developed inside the classroom. These activities were developed individually and were valued according to some preestablished principles (rubric).

Test 1 (multiple option) consisted of 30 items having between three and five possible answers from which to select. Test 2 (paired) consisted of four blocks of questions, with each block containing 8 to 12 questions, for a total of 41 questions. The structure of the two tests assumes the implications and reasoning presented in the literature in this respect (Berrios, Rojas, & Cartaya, 2005). Both tests were implemented in the Virtual Campus of the course through the data processing platform Moodle (Dougiamas), although currently it is feasible to apply them in other similar platforms (Tuparov & Dureva-Tuparova, 2008). The Moodle platform allows evaluations to take place virtually inside (our case) or outside the classroom, and evaluation is done using the previous test program. As a result of the process, the system generates an output file in Word, Excel, or RTF, thus allowing processing.

Table 4. Nomenclature of the Study Variables.

Nomenclature	Meaning	Value
VAR01	Group to which belong the students	1M = 1, 2M= 2, 3T = 3, and 4T =4
VAR02	Test 1	From 0 to 10[a]
VAR03	Test 2	From 0 to 10[a]
VAR04	Final score	From 0 to 10[a]

[a] Accuracy of two decimals.

Tests were defined based on the following criteria and data processing adjustments, which help to standardize their application (regulations were provided to the student body prior to administration of evaluations):

1. Maximum time of completion: 1 min for each question.
2. For each item, the last answer given is considered for scoring.
3. Number of attempts per question: Unlimited.
4. Penalty in each question: The proportional part of the question value divided by the number of possible answers.
5. Value of each question: All questions have the same proportional weight inside the global test.

The tests were also proposed to evaluate the different knowledge levels achieved by students, based on the Taxonomy of Bloom (Van Niekerk & Von Solms, 2009). Table 3 summarizes the subdivision of knowledge levels evaluated, including the number of questions for each one of them.

For the analysis in the first part of the statistical study, four different variables were used. Table 4 shows the codes and meanings assigned to these variables. With the criteria given earlier and the variables to analyze, the data processing program SPSS V17 for Windows was utilized, for the purpose of obtaining the general descriptive statistical parameters of each variable, in a separated form, and thus to understand and distinguish them. The studied parameters were as follows.

Table 5. Results of the Descriptive Statistics for the Study Variables.

				4T					
Group		1M, 2M, and 3T		Test 1			Test 2		
Parameter		VAR01	VAR04	VAR01	VAR02	VAR04	VAR01	VAR03	VAR04
No. Valid[a]		247	247	32	32	32	39	39	39
No. Lost[a]		0	0	215	215	215	208	208	208
M		2.55	5.74	4.00	7.86	6.81	4.00	8.95	6.57
SE of the mean		0.07	0.14	0	0.18	0.26	0	0.29	0.22
Median		3.00	6.30	4.00	7.81	7.00	4.00	10.00	7.00
Mode		4.00	0.00	4.00	7.80	7.00	4.00	10.00	7.00
SD		1.13	2.27	0.00	1.04	1.45	0.00	1.80	1.38
Variance		1.28	5.17	0.00	1.08	2.09	0.00	3.24	1.91
Asymmetry		−0.06	−1.53		−0.86	−3.57		−2.44	−2.91
SE of asymmetry		0.15	0.15	0.41	0.41	0.41	0.38	0.38	0.38
Kurtosis		−1.39	1.57		0.25	16.20		6.10	13.03
SE of kurtosis		0.31	0.31	0.81	0.81	0.81	0.74	0.74	0.74
Amplitude		3.00	8.68	0.00	3.75	8.50	0.00	7.50	8.50
Minimum		1.00	0.00	4.00	5.39	0.00	4.00	2.50	0.00
Maximum		4.00	8.68	4.00	9.14	8.50	4.00	10.00	8.50
Sum		631.00	1,417.39	128.00	251.40	217.80	156.00	349.09	256.30
Percentiles	0			4			4		
	10	1	0	4	6.17	5.30	4	7	5.20
	20	1	5	4	7.21	6.88	4	8	6
	30	2	5.44	4	7.38	7	4	9	6.10
	40	2	5.95	4	7.72	7	4	9.50	7
	50	3	6.30	4	7.81	7	4	10	7
	60	3	6.87	4	8.34	7	4	10	7
	70	4	7	4	8.64	7.01	4	10	7
	80	4	7.48	4	8.81	7.50	4	10	7.10
	90	4	7.80	4	9.09	8	4	10	8
	100	4	8.68	4	9.14	8.50	4	10	8.50

[a]Number of included students.

1. Central tendency measures (mean, median, mode, and sum);
2. Dispersion measures (standard deviation, variance, amplitude, minimum, maximum and error of mean), sampling distribution (asymmetry and kurtosis), and finally the percentile values.

Table 5 presents the general results obtained for the four analyzed variables regarding their general statistical description.

Discussion

The Analyzed Variables

With respect to the measures of central tendency, one can say that the final score average of the study groups are located in the range of 5.5 to 7.0 (high score possible 10). Groups 1M, 2M, and 3T have an average score, but not Group 4T (of study). The study tests (Tests 1 and 2) applied to Group 4T showed average values over the values above (close to 8.0 for Test 1 and 9.0 for Test 2; high score possible for both cases of 10). From the above-mentioned values, it can be said that Group 4T performs better than the other reference groups, and that the tests discussed in this article do not represent difficulty in resolution. Therefore, their use as a teaching tool has a manageable difficulty in this course (test appropriate to the content of the subject to value; see Table 5 and Figure 1). Finally, the students who took Tests 1 and 2 were about the same in number, thus improving the interpretation and correlation between variables.

With regard to measures of dispersion, as shown in Table 5 and Figure 2, the standard deviation is always less (for the final score, and for both tests studied in this work) than the average score end of the reference groups (1M, 2M, and 3T), with a difference of about 0.5 unit and becoming similar when the coefficients of variation of the test study and final average score groups study are compared. Therefore, this indicates that both the test study, and the behavior of the results of the control groups are substantially the same, and

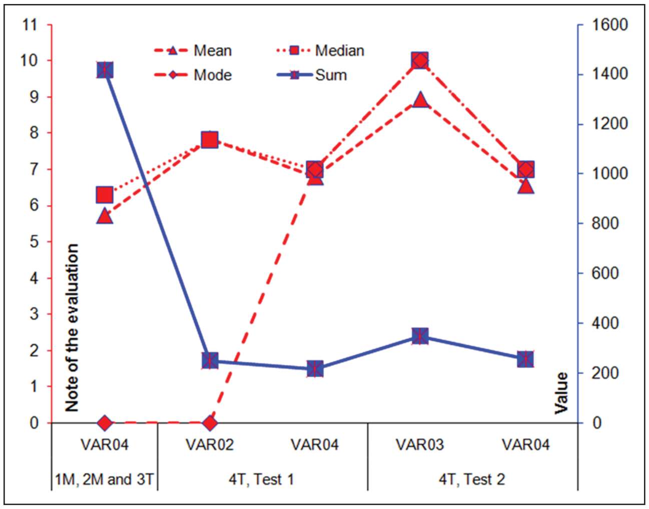

Figure 1. Central tendency measures.

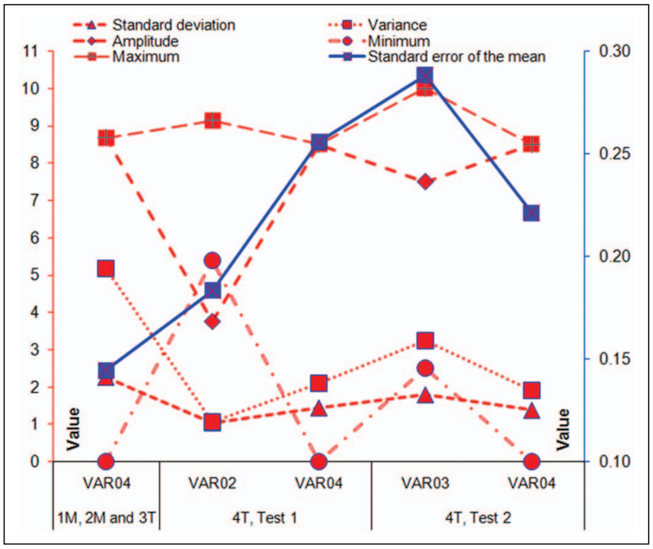

Figure 2. Dispersion measures.

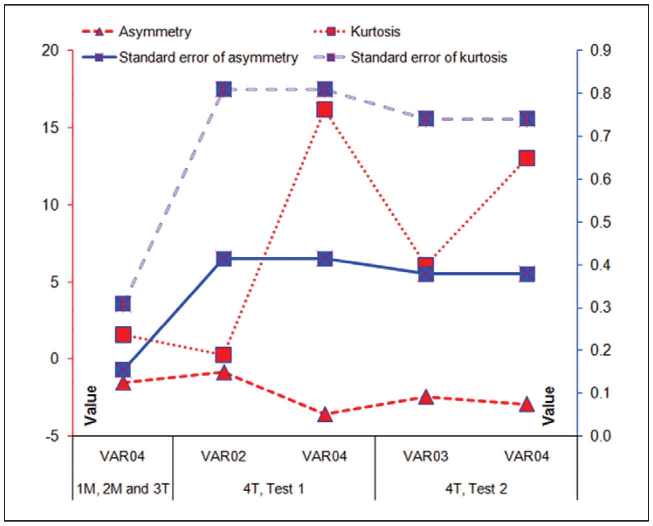

Figure 3. Form of the curve distribution.

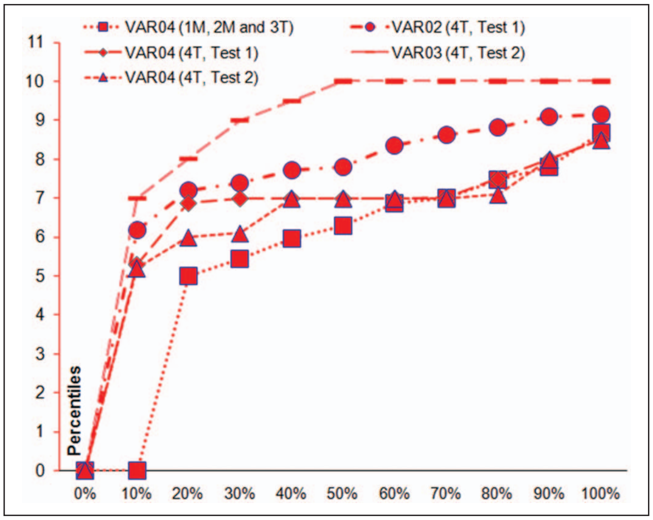

Figure 4. Distribution of the notes.

the study variables are related among themselves, anticipating the absence of other variables.

Moreover, it may be observed that the amplitude of the scores achieved in each of the evaluations indicate that the tests applied "focus" better on student scores (amplitude of 3.5 for Test 1 and 7.5 for Test 2, while for the control groups and final score study group the amplitude reached between 8.5 and 9.5 point difference).

With respect to the shape of the distribution curve having the different variables of study, it can be seen that the elevation of the distribution is more pronounced for the test case study (comparing the average score at the end of the course achieved for both the control and study groups, between 12 and 15 times higher; see Table 5 and Figure 3). Similarly, it can be said that the data distributions for all variables have unilateral values, extending into the negative zone (left branch distributions are larger).

Finally, with reference to the distribution of the scores achieved by students, for each of the study variables (see Table 5 and Figure 4), we can say that the average value of the final score for the control groups is linear, incremental, and positive, much like the notes of Test 1 (similar slope, but increased at the beginning), while in the case of Test 2, it becomes constant curvature from 50%. This may help provide an understanding of how to distribute the scores and compare different evaluation techniques.

In conclusion, we can highlight two general ideas from the comparative statistics. First, the mean of the score of the students (VAR04 = final score) who took the test (VR02 and VR03) is higher (Group 4T) than that of students who did not take the test (Groups 1M, 2M, and 3T). Second, the variance of the results is smaller for the group (4T) that took the test (VR02 and VR03).

Psychometric Analysis of the Items

Psychometric analysis is a mathematical procedure that applies statistical principles to determine the suitability of the proposed questions based on the responses and their individual relationship with the rest of the answers, thereby detecting whether the proposed questions are appropriate to assess the level of knowledge, degree of difficulty, and degree of discrimination between high and low conceptual skills (Heck & Van Gastel, 2006; Revuelta, Ximénez, & Olea, 2003).

From the results of the multiple-option and the paired tests, as previously discussed, some parameters were extracted and utilized. These are defined and analyzed in Tables 6 and 7 where the processed data of the surveys are presented in a manner that permits the analysis and evaluation of the performance of each question, taking into account the global evaluation of the sample. The statistical parameters utilized in these tables were determined with the evaluation of the classical theory of tests (Batrinca & Raicu, 2010; General Public License GNU, 2010). The theory behind the analysis chosen to calibrate questionnaires or assess psychometric properties are not presented in this work, as on one hand, the system used is the existing in college, and on the other hand, his theoretical justification are in the tool and can be found on the WEB (Dougiamas).

The first parameter presented in the Tables 6 and 7, for the analysis of the tests, is the Facility Index (FI; % correct), which is defined as the mean value of how easy or difficult an item is, with regard to the rest of the questions inside the same analysis group (test). This parameter is determined with the following equation:

$$\mathrm{FI} = \frac{X_{\mathrm{mean}}}{X_{\mathrm{max}}}, \tag{1}$$

where X_{mean} is the mean value from all values obtained for the total users who did every item and X_{max} is the maximum value obtained for that item.

If the questions could be distributed in dichotomous categories (*correct*/*incorrect*), this parameter would coincide with the percentage of students who responded to the questions correctly.

In our study, and considering Figure 5, most of the questions in Test 1 are concentrated on the band from 70% to 90% of FI, while in Test 2, they are located in a band from 85% to 90%. From these results, it is deduced that the questions or blocks of questions located out of both extremes of previous bands should be eliminated in future editions of the test because they are trivial (FI very low) or they are of a high difficulty level (FI very high). In either possibility, these questions should not be utilized as criteria to discern an educational evaluation, because they are not useful as evaluation criteria. The graph in Figure 5 shows the areas discussed.

Another possible alternative in deciding which questions or blocks of questions could be eliminated from a test is to verify that the questions are correctly defined, not including errors in their formulation and complying with basic criteria of logic. To accomplish this task, an exhaustive review of the editing, structure, logic, and coherence of questions must be done before using them again in another evaluation.

The second parameter evaluated in this work is the standard deviation (*SD*), which indicates the dispersion of the response in relation to the answers given by the entire population analyzed. As a comment to this parameter, it can be said that in the event that all students respond equally to a specific question (item), the value obtained for *SD* would be zero.

SD is obtained with the statistical standard deviation of the sample (classical analytic statistical), or if not, with the mark of class (relation obtained/maximum) for each specific item.

In our case, and considering Figure 6, this parameter can be utilized as a criterion of detection to verify the knowledge acquisition by part of the student body in a determined concept or item. This knowledge contributed by *SD* should not be seen as particular or individual; the correct interpretation is from a perspective that is most general and uniform for all the members (collective general knowledge of the theme).

In Test 1, the questions that surpass the upper band of the established criterion (in this case, it could be set as an *SD* close to 0.30) are questions with thematic content advisable to be reviewed again in the classroom to guarantee some minimum content learned by all students.

For Test 2, there is a great divergence between the two clearly defined groups of *SD*. Thus, the form in which the questions have been grouped (paired questions) should be changed. The four blocks of items should be centered, improving the verification uniformity of the acquired knowledge. The graph in Figure 6 shows the area discussed.

Another interesting parameter for the analysis of test results is the Discrimination Index (DI), which provides an approximate indicator of each item (question) or analyzed response (separately) on its performance with regard to the answer with a smaller performance level. This way, it allows one to deduce between high punctuation with respect to global punctuation, and a less-expert user with respect to the experienced.

This parameter is obtained by dividing the student group analyzed by thirds, keeping in mind its scoring with reference to the global questionnaire. Below, for the superior and inferior groups the average punctuation from the analyzed item is obtained (continuing the performance order of up downward); finally, from the previous value is subtracted the average of the punctuation. The mathematical expression is as follows:

Table 6. Details of Test Parameters for the Study: Multiple-Choice Test, Topic 9.

Question No.	Possible answer	Possible value for each individual answer	No. of times responded/Total no. responded for question	FI (%)	*SD*	DI	DC
1	1	1	30/32	82	0.34	0.96	0.78
	2	−0.33	0/32				
	3	−0.33	0/32				
2	1	−0.33	0/32	92	0.25	0.83	0.32
	2	−0.33	0/32				
	3	1	30/32				
3	1	−0.33	0/32	63	0.29	0.58	0.22
	2	1	30/32				
	3	−0.33	2/32				
4	1	−0.33	0/32	86	0.34	1	0.75
	2	−0.33	0/32				
	3	1	28/32				
5	1	−0.33	2/32	76	0.35	0.58	0.02
	2	−0.33	0/32				
	3	1	30/32				
6	1	1	30/32	90	0.16	0.63	−0.17
	2	−0.33	0/32				
	3	−0.33	0/32				
7	1	1	30/32	84	0.21	0.63	−0.01
	2	−0.33	0/32				
	3	−0.33	0/32				
8	1	−0.33	0/32	82	0.24	0.71	0.24
	2	1	32/32				
	3	−0.33	0/32				
9	1	−0.33	0/32	76	0.26	0.79	0.60
	2	−0.33	0/32				
	3	1	30/32				
10	1	−0.33	0/32	71	0.29	0.63	0.17
	2	−0.33	0/32				
	3	1	32/32				
11	1	−0.25	0/32	63	0.28	0.53	0.07
	2	−0.25	0/32				
	3	−0.25	0/32				
	4	1	30/32				
12	1	−0.33	2/32	84	0.27	0.75	0.31
	2	1	30/32				
	3	−0.33	0/32				
13	1	1	28/32	73	0.36	0.71	0.27
	2	−0.33	0/32				
	3	−0.33	0/32				
14	1	−0.25	2/32	81	0.33	0.82	0.42
	2	−0.25	0/32				
	3	−0.25	0/32				
	4	1	30/32				
15	1	−0.33	0/32	80	0.30	0.71	0.28
	2	−0.33	2/32				
	3	1	30/32				
16	1	−0.25	0/32	82	0.25	0.70	0.32
	2	−0.25	0/32				
	3	−0.25	0/32				
	4	1	30/32				

(continued)

Table 6. (continued)

Question No.	Possible answer	Possible value for each individual answer	No. of times responded/Total no. responded for question	FI (%)	*SD*	DI	DC
17	1	1	26/32	55	0.35	0.42	0.01
	2	−0.33	4/32				
	3		0/32				
18	1	−0.25	0/32	71	0.31	0.81	0.57
	2	−0.25	0/32				
	3	−0.25	0/32				
	4	1	30/32				
19	1	−0.33	0/32	96	0.11	0.71	−0.15
	2	1	32/32				
	3	−0.33	0/32				
20	1	1	30/32	73	0.32	0.72	0.45
	2	−0.33	0/32				
	3	−0.33	0/32				
21	1	−0.33	0/32	88	0.24	0.70	0.15
	2	1	30/32				
	3	−0.33	0/32				
22	1	−0.25	0/32	75	0.27	0.72	0.56
	2	−0.25	0/32				
	3	−0.25	0/32				
	4	1	30/32				
23	1	1	32/32	86	0.27	0.83	0.47
	2	−0.33	0/32				
	3	−0.33	0/32				
24	1	−0.33	0/32	78	0.31	0.83	0.61
	2	−0.33	0/32				
	3	1	30/32				
25	1	1	30/32	84	0.29	0.83	0.59
	2	−0.33	0/32				
	3	−0.33	0/32				
26	1	−0.33	0/32	80	0.27	0.83	0.68
	2	−0.33	0/32				
	3	1	30/32				
27	1	−0.25	0/32	66	0.32	0.69	0.36
	2	−0.25	0/32				
	3	1	30/32				
	4	−0.25	0/32				
28	1	−0.20	0/32	92	0.25	0.85	0.37
	2	1	30/32				
	3	−0.20	0/32				
	4	−0.20	0/32				
	5	−0.20	0/32				
29	1	−0.25	0/32	69	0.37	0.56	0.11
	2	−0.25	0/32				
	3	−0.25	0/32				
	4	1	28/32				
30	1	−0.25	0/32	76	0.36	0.88	0.55
	2	1	30/32				
	3	−0.25	0/32				
	4	−0.25	0/32				

Note. Possible answer: order in which each possible answer is presented; Possible value for each individual answer: reduction of the punctuation (incorrect response) and increase of the punctuation (correct response); No. of times responded/Total no. responded for question: number of times that this question is answered with reference to the total of possible answers of the test.

Table 7. Details of Test Parameters for the Study: Paired Test, Topic 10.

No. of the question block	Possible value for each individual answer	No. of times responded/Total no. responded for question	FI (%)	*SD*	DI	DC
1	1	38/39	94	0.17	0.98	0.87
		38/39				
		38/39				
		38/39				
		38/39				
	0	1/39				
		1/39				
		1/39				
		1/39				
		1/39				
2	1	39/39	90	0.17	0.90	0.77
		39/39				
		38/39				
		36/39				
		36/39				
	0	1/39				
		2/39				
		2/39				
3	1	37/39	89	0.27	0.96	0.86
		38/39				
		36/39				
		37/39				
		37/39				
		38/39				
	0	1/39				
		1/39				
		1/39				
		1/39				
		1/39				
4	1	37/39	85	0.245	0.89	0.78
		35/39				
		38/39				
		36/39				
		38/39				
	0	3/39				
		3/39				
		1/39				
		1/39				
		1/39				
		1/39				
		1/39				

Note. Possible value for each individual answer: reduction of the punctuation (incorrect response) and increase of the punctuation (correct response); No. of times responded/Total no. responded for question: number of times that this question is answered with reference to the total of possible answers of the test.

$$\mathrm{DI} = \frac{\left(X_{top} - X_{bottom}\right)}{N}, \tag{2}$$

where X_{top} is the sum of the reached fraction (obtained/maximum) for this item, for a third of students with higher qualifications in the whole questionnaire; this is the number of correct answers in this group; and X_{bottom} is the analog sum for the students located in the lower third of the questionnaire.

This parameter has values in the range of +1 to −1. Its meaning should be interpreted as follows: When DI is getting greater than 0.0, more low-performance students have been assumed to be better in this item than students with

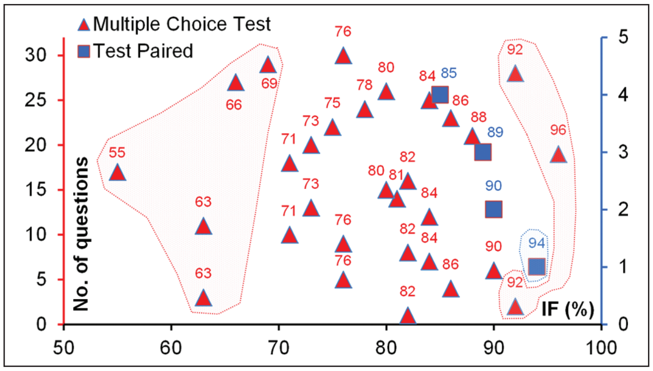

Figure 5. Index results for the tests facility for study.

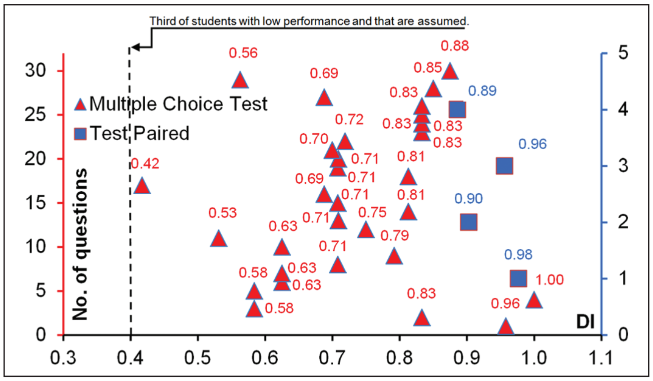

Figure 7. Discrimination Index results for the tests study.

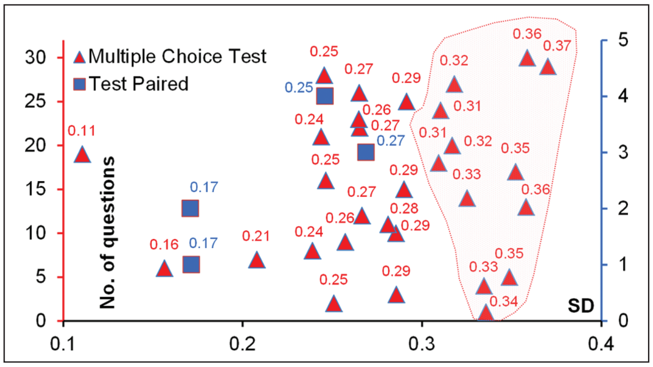

Figure 6. Standard deviation results for the tests study.

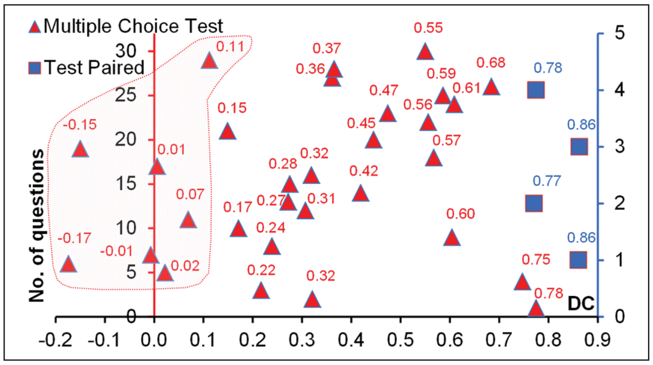

Figure 8. Discrimination coefficient results for the tests study.

higher performance. Therefore, these items, as questions for evaluation, should be eliminated for being inadequate. In fact, these items reduce the global score precision of the test.

In our work (Figure 7), and with the aim of validating an evaluative questionnaire, it will be necessary to eliminate the questions in Test 1 that have a DI lower than 0.4 because these are located in the third of students with low performance and having assumed knowledge assessed. It is important to note that, in this case, these questions are not badly designed, but they are not necessary for evaluation because of their simplicity. The graph in Figure 7 shows the border discussed.

In Test 2, the concepts before established for Test 1 are applicable, thus completing this questionnaire, with exceeded reliability for future applications. Therefore, it is necessary to adjust the test for application in new practices.

The last statistical parameter analyzed in this work is the Discrimination Coefficient (DC), which is considered another parameter of measure to achieve the separation of adequate items and low-performance items from the learning evaluation.

DC is a coefficient of correlation among the scores of each particular item with respect to the complete questionnaire. Its mathematical expression is as follows:

$$\mathrm{DC} = \frac{\sum(xy)}{N \bullet S_x \bullet S_y}, \quad (3)$$

where $\Sigma(xy)$ is the products' summation of the deviations for the samples marks of items, with reference to the total survey or test, N is the number of answers obtained for a question or item, S_x is the standard deviation value of the results for the fraction of the question, and S_y is the standard deviation value of the results of the total questionnaire.

As in the previous parameter (DI), DC can obtain a range of values from +1 to −1. Positive values indicate items that discriminate right questions, while indices with negative values are items that are answered by low-performing students. This means that items with a negative DC are answered incorrectly by students, which penalizes the majority of students. Therefore, these topics or test questions must be removed.

The advantage of DC with respect to DI is that the former utilizes the entire population of the analysis group to obtain information for its decision, and not just the extreme upper and lower thirds as DI does. Consequently, DC can be considered more sensitive in detecting the performance of the items or questions. In our case, as shown in Figure 8, the detection of the ineligible questions to be considered in

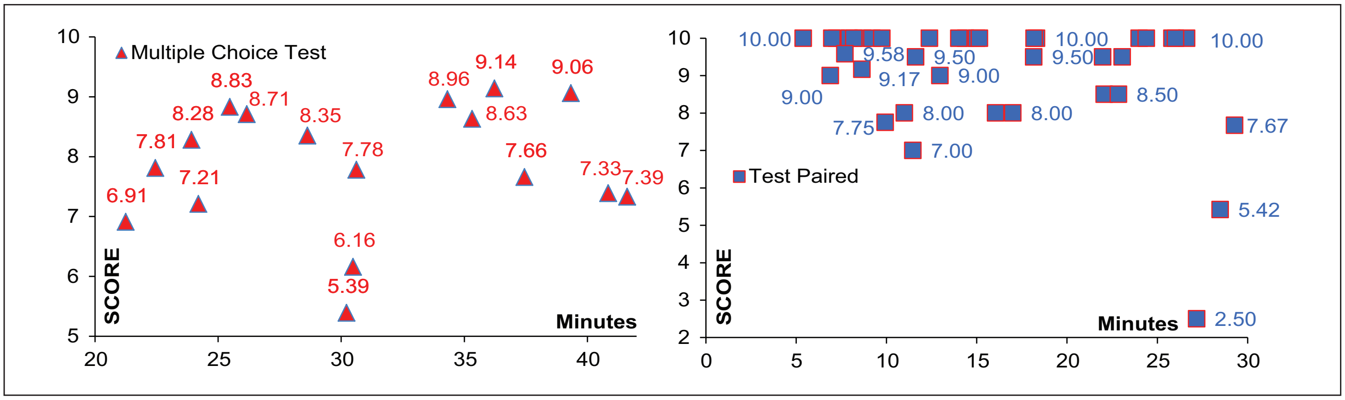

Figure 9. Test resolution study versus time resolution.

future versions of tests is more evident with DC than with DI. The graphs of Figures 7 and 8 show the comments.

For Test 1, besides Question 17, which was detected by DI, Questions 5, 6, 7, 11, 19, and possibly 29 also show serious problems in their resolution by part of the students. For Test 2, the only difference in the use of DI with regard to the values reported using DC is the value reached in its scale, as well as its higher proximity to the nil value. However, DI and DC describe similar order and relation.

To finish, although the following comments are out of the scope of the statistical analysis of the test in this work, Figure 9 shows our case average time employed in the resolution of the test with reference to the average grade reached by students. In general conditions, and for the case of Test 1, there are scores with high values unrelated to the time spent in the resolution. This fact could be used to detect concepts used in the learning, such as bright or effective students. However, students with low scores, who recognize their knowledge deficiency, decline to use adequately all the available time to resolve the questionnaire. In the case of Test 2, the students who achieve high or medium scores do not utilize the total available time (up to 27 min), whereas students with low scores use such time. It is evident that in this test, the resolution time should be adjusted downward, to better adapt its use and evaluation.

Monitoring of the Process and Result of the Improvement

To verify the adaptations, modifications, and replacement proposed in both tests, we performed a second evaluation on students of the following year. For this occasion, the thematic content and teacher were the same, but the students were different. In the comparative analysis of "pre" and "post" test, we observed an improvement in the control parameters.

To obtain control parameters representing the test study, we obtained the average of the results of each index calculated before (FI, *SD*, DI, and DC), summing the individual value of each question and then dividing by the number of test questions. These parameters measured in global terms whether a test is easier than the other or if the results are more uniform or dispersed.

The values of the control parameters (initial mean) of the test of this study were associated with the name of "pretest," that is, in reference to the proposed initial tests before identifying potential improper or incorrect questions. The values thus obtained are as follows:

Pretest:

For FI:

Test 1: between 70% and 90% (average 74.47%); Test 2: between 85% and 90% (average 89.5%).

For *SD*:

Test 1: average = 0.28; Test 2: average = 0.21.

For DI:

Test 1: average = 0.73; Test 2: average = 0.93.

For DC:

Test 1: average = 0.33; Test 2: average = 0.82.

In the second evaluation, the following questions were revised (in accordance with the previous study): For Test 1: trivial questions (3, 11, 17, 27, 29), hard questions (2, 19, 28). For Test 2: trivial questions (1). For Test 1: having high *SD* (1, 4, 5, 13, 14, 17, 18, 20, 24, 27, 29, 30). For Test 1: presenting a very low DI (17). For Test 1: presenting a very low DC (5, 6, 7, 11, 17, 19, 29). Thus, a new test was verified, with the appropriate calibration undertaken, for use with a new generation of students in the same educational institution (next school year).

As a result of this second evaluation, the average control parameters were assessed again and associated with the term *posttest*. The results obtained were as follows:

Posttest:

For FI:

Test 1: between 69% and 81% (average 72.1%); Test 2: between 75% and 84% (average 73.5%).

For *SD*:

Test 1: average = 0.25; Test 2: average = 0.22.

For DI:

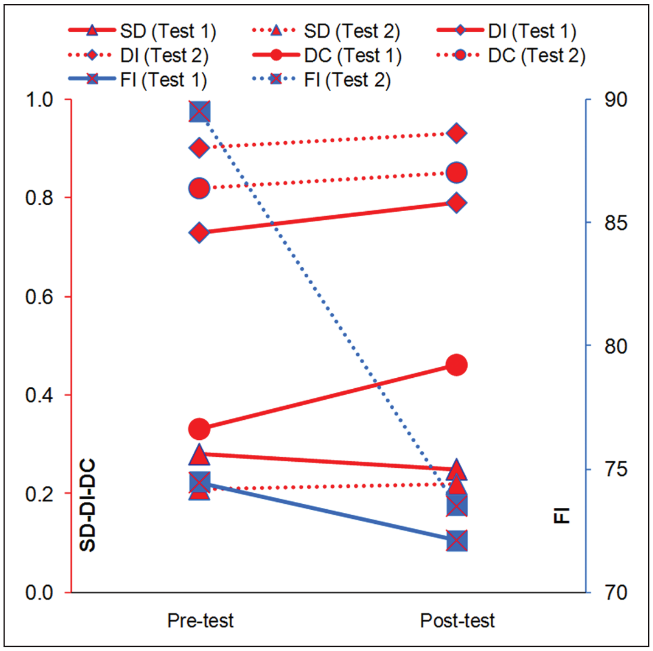

Figure 10. Improvement checking in the test analyzed.

Test 1: average = 0.79; Test 2: average = 0.93.

For DC:

Test 1: average = 0.46; Test 2: average = 0.85.

Finally, comparing the results pretest and posttest (see Figure 10), one can say that the average FI parameters facilitate the resolution of the test (negative slope of the lines between pretest and posttest). However, lower values are reported in the case of *SD* (greater uniformity of response in the test). Finally, for the case of DI and DC, values are reported at the initial upper level (positive slope of the lines between test), indicating that these tests are more robust and useful as a tool to assess student knowledge.

Conclusion

The final general comments are as follows:

At the Moodle platform, the tool "Questionnaires" gives faculty the possibility to implement active learning and self-learning experiences for educational purposes. It is also a simple-use instrument that is suitable for evaluating the knowledge level reached by students.

The use of the available questionnaires on this platform is a big and versatile tool, with applications in educational aspects, such as self-learning and learning evaluation, and as a criterion for particular adaptation in teaching.

This tool allows the promotion of learning activities outside the classroom, reduction in evaluation times (especially in big groups of students), and detection of specific or particular needs of a student or group of students.

The implementation of this tool requires extra work by the teacher at the beginning of its use. This initial effort is compensated with the satisfaction that comes from meeting the predicted educational expectations, improvements in the educational level reached, and the acceptance of its use by students.

The specific final comments of this work are as follows:

The processed information obtained in tests can contribute "extra information" that allows adapting the entire teaching process in a better form.

The FI permits discernment among the difficulty levels of the questions established in a test, so it can be used as a criterion to select questions, and thus to guarantee the adaptation of each of them, or in lack of that, a scrupulous review of its logic.

The *SD* permits the detection of knowledge acquisition by students. This parameter has a general and uniform character for all the members of the group (general collective knowledge of the theme). Thus, it contributes criteria of what is or what is not learned by students.

The DI allows one to detect those questions that should be eliminated in tests because they are inadequate for evaluation. This way, the precision of the global score of the test can be improved. It is important to note that these questions are not badly designed, but they are not necessary to evaluate because of their simplicity.

The DC permits one to obtain a parameter with detection of ineligible questions in a test. This is a more sensitive parameter than DI, as it can be used to select with success those items more adequate for the knowledge evaluation of students.

The control and analysis of the time used in the evaluation test can contribute with adjustments and additional information on the entire evaluation process.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research and/or authorship of this article: The authors thank Project S0117-CTTP-UPC for the financing and the group of investigation GICITED-UPC.

References

Batrinca, G., & Raicu, G. (2010). *Considerations about effectiveness and limits of computer based training in maritime industry*. In Proceedings of the 3rd International Conference on Maritime and Naval Science and Engineering, Romania (pp. 15-20).

Berrios, G., Rojas, C., & Cartaya, N. (2005). Effect of the number of options on the quality of EST reading comprehension multiple-choice exams. *Paradígma*, *26*(1), 89-116.

Blanco, M., Estela, M., Ginovart, M., & Saà J. (2009). Computer assisted assessment through Moodle quizzes for calculus in an engineering undergraduate course. *CIEAEM 61 Quaderni di Ricerca in Didattica Matematica*, *19*, 78-83.

Blanco, M., & Ginovart, M. (2009). *Creating Moodle quizzes for the subjects of mathematics and statistics corresponding to the first years in engineering studies*. EDULEARN09 Proceedings CD of the International Conference on Education and New Learning Technologies. IATED, 1984-1993.

Blanco, M., & Ginovart, M. (2010). *Moodle quizzes for assessing statistical topics in engineering studies*. Joint International IGIP-SEFI Annual Conference, 2010, Trnava, Slovakia (pp. 1-4).

Bowles, T. (2008). Self-rated estimates of multiple intelligences based on approaches to learning. *Australian Journal of Educational & Developmental Psychology*, *8*, 15-26.

Britt, K., McCall, L., Austin, D., & Piterman, L. (2007). A psychometric evaluation of the learning styles questionnaire: 40-item version. *British Journal of Educational Technology*, *38*, 23-32.

Burden, P. (2008). ELT teacher views on the appropriateness for teacher development of end of semester student evaluation of teaching in a Japanese context. *System Science Direct Elsevier*, *36*, 478-491.

Dougiamas, M. (1998). *Moodle*. Retrieved from http://moodle.org/

Escuela Politécnica Superior de Edificación de Barcelona [School of Building Construction of Barcelona]. (2009). *Construciones Arquitectónicas II. Universidad Politécnica de Cataluña. Memoria 2007-2008*. Retrieved from http://www.tifeb.upc.edu/files/escola/memoria/memoria07-08/curso.pdf

Esquivel, J. M. (2000). *El diseño de las pruebas para medir logro académico: ¿Referencia a normas o criterios?* [The design of the tests to measure academic achievement: Reference to standards or criteria?]. (P. Ravela, Ed.). PREAL/GRADE. Retrieved from http://www.oei.es/calidad2/grade.PDF

Fančovičová, J., & Prokop, P. (2010). Development and initial psychometric assessment of the plant attitude questionnaire. *Journal of Science Education and Technology*, *19*, 415-421.

Garfield, J. B. (2003). Assessing statistical reasoning. *Statistics Education Research Journal*, *2*(1), 22-38.

General Public License GNU. (2010). *Moodle for teachers, trainers and administrators*. Retrieved from http://moodle.org/

Goldston, M. J., Dantzler, J., Day, J., & Webb, B. (2012). A psychometric approach to the development of a 5E lesson plan scoring instrument for inquiry-based teaching. *Journal of Science Teacher Education*. Advance online publication. doi:10.1007/s10972-012-9327-7

Gómez-Soberón, J. M. (2009). *Adaptación de las asignaturas de construcción del Departamento de Construcciones Arquitectónicas II al nuevo concepto de los créditos europeos de Educación Superior y del campus virtual Atenea* [Adapting construction subjects (Department of Architectural II) to the new concept of European Higher Education Credit and Athena virtual campus]. Barcelona, Spain: Creative Commons. Retrieved from http://hdl.handle.net/2117/8369

Gómez-Soberón, J. M., Bosch-González, M., Marín-i-Gordi, O., Gómez-Soberón, M. C., & Gómez-Soberón, L. A. (2010). *Statistical analysis and use of questionnaires for evaluation of the knowledge at university. A practical case*. La Habana, Cuba: Ministerio de Educación Superior y Universidad de la República de Cuba. Retrieved from http://hdl.handle.net/2117/9672

Gómez-Soberón, J. M., & Gómez-Soberón, M. C. (2007). Aplicación de unatécnica de mejora docente. [Application of a technique of improving teaching]. Jornadasde Enseñanza del Hormigón Estructural. [Conference on Teaching StructuralConcrete]. Madrid, Spain: Asociación Científica del Hormigón Estructural [Structural ConcreteAssociation]. Retrieved from http://hdl.handle.net/2117/2848

Gómez-Soberón, J. M., Gómez-Soberón, M. C., & Gómez-Soberón, L. A. (2009). *Active learning and autonomous learning: An educational experience*. Barcelona, Spain: ALE-2009. The Learning Experience. Retrieved from http://hdl.handle.net/2099/7809

Heck, A., & Van Gastel, L. (2006). Mathematics on the threshold. *International Journal of Mathematical Education in Science and Technology*, *37*, 925-945.

Hutchison, D. (2009). Designing your sample efficiently: Clustering effects in education surveys. *Educational Research*, *51*, 109-126.

Jaeger, R. M. (1998). Evaluating the psychometric qualities of the national board for professional teaching standards' assessments: A methodological accounting. *Journal of Personnel Evaluation in Education*, *12*, 189-210.

Mowder, B. a., & Shamah, R. (2011). Parent Behavior Importance Questionnaire–Revised: Scale development and psychometric characteristics. *Journal of Child and Family Studies*, *20*, 295-302. doi:10.1007/s10826-010-9392-5

Murayama, K., Zhou, M., & Nesbit, J. C. (2009). A cross-cultural examination of the psychometric properties of responses to the achievement goal questionnaire. *Educational and Psychological Measurement*, *69*, 266-286. doi:10.1177/0013164408322017

Myford, C. M., & Engelhard, G., Jr. (2001). Examining the psychometric quality of the national board for professional teaching standards early childhood/generalist assessment system. *Journal of Personnel Evaluation in Education*, *15*, 253-285.

Nulty, D. D. (2008). The adequacy of response rates to online and paper surveys: What can be done? *Assessment & Evaluation in Higher Education*, *33*, 301-314.

Ravela, P. (Ed.). (2000). *Los próximos pasos: ¿hacia dónde y cómo avanzar en la evaluación de aprendizajes en América Latina?* [Next steps: where and how to advance in the assessment of learning in Latin America?]. PREAL/GRADE. Retrieved from http://www.oei.es/calidad2/grade.PDF

Revuelta, J., Ximénez, M. C., & Olea, J. (2003). Psychometric and psychological effects of item selection and review on computerized testing. *Educational and Psychological Measurement*, *63*, 791-808

Rivero, N., Martínez-Pampliega, A., & Olson, D. O. (2010). Spanish adaptation of the FACES IV questionnaire: Psychometric characteristics. *The Family Journal: Counseling and Therapy for Couples and Families*, *18*, 288-296. doi:10.1177/1066480710372084

Romero, C., Ventura, S., & Salcines, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computer & Education*, *51*, 368-384.

Shepard, L. A. (2006). *La evaluación en el aula. Textos de evaluación* (Instituto Nacional para la Evaluación de la Educación, Ed.) [The assessment in the classroom: Texts evaluation. National Institute for Educational Evaluation (Ed.)]. Distrito Federal, Mexico.

Tiana, A. (1997). *Tratamiento y usos de la información en evaluación. Calidad y Equidad en la Educación* [Treatment and use of assessment information. Quality and equity in education]. Iberoamericano Organización de Estados, Programa Evaluación de la Calidad de la Educación. Retrieved from http://www.oei.es/calidad2/tiana.htm

Tuparov, G., & Dureva-Tuparova, D. (2008). *On-line testing implementation in open source e-learning environments*. In International Scientific Conference Computer Science (pp. 875-880).

Van Niekerk, J., & Von Solms, R. (2009). *Using Bloom's taxonomy for information security education*. Education and Technology for a Better World: 9th IFIP World Conference on Computers in Education (WCCE), Bento Gonçalves, RS, Brazil (pp. 27-31). Retrieved from http://www.wcce2009.org/proceedings/papers/WISE6_Niekerk.pdf

Veiga, A., & Amaral, A. (2009). Survey on the implementation of the Bologna process in Portugal. *Higher Education: The International Journal of Higher Education and Educational Planning*, *57*(1), 57-69.

Zhang, D. (2011). The psychometric evaluation of a three-dimension elementary science attitude survey. *Journal of Science Teacher Education-Elementary Science Education*, *22*, 591-612.

Author Biographies

J. Manuel Gómez-Soberón is a civil engineer with PhD in construction from Universidad Politécnica de Cataluña (UPC; 2002). He is a professor in the School of Construction at the UPC in Barcelona, Spain, and a researcher with interests in concrete recycling, sustainable construction, recycling of construction materials, active learning, assessment in teaching, and innovative use of Information and communication technologies (ICT) in teaching.

M. Consolación Gómez-Soberón is a civil engineer with PhD in structural engineering from the UPC (2002). She is a professor in the School of Civil Engineers UAN in Mexico DF and a researcher with interests in vulnerability of bridges, seismic structural behavior, behavior of soil-structure, and building energy sinks.

Ramón Corral-Higuera is a civil engineer with PhD in science (materials) from Centro de Investigación y de Estudios Avanzados del Instituto PolitécnicoNacional (CIEA IPN; Center for Research and Advanced Studies of the National PolytechnicInstitute), Unit of Chihuahua, Mexico (2010). He is a professor in the School of Civil Engineers Universidad Autónoma de Sinaloa (UAS) in Los Mochis Sinaloa, Mexico, and a researcher with interests in recycled concrete, durability of concrete, and sustainable construction.

S. Paola Arredondo-Rea is a civil engineer with PhD in science (materials) in CIEA IPN, Unit of Chihuahua, Mexico (2010). She is a professor in the School of Civil Engineers UAS in Los Mochis Sinaloa, Mexico, and a researcher with interests in physical–chemical and microstructural sustainable materials (recycled concrete products of industrial processes).

J. Luis Almaral-Sánchez is a civil engineer and PhD of science (materials) in CIEA IPN, Unit of Queretaro, Mexico. He is a professor in the School of Civil Engineers UAS in Los Nochis Sinaloa, Mexico, and a researcher with interests in steel corrosion, organic–inorganic hybrid materials, zeolites, nanotechnology, and sustainable materials.

F. Guadalupe Cabrera-Covarrubias is a civil engineer and a PhD candidate in civil engineering from the UPC in Barcelona Spain. Her research interests are mortars recycled, applications in construction, and recycled concrete.