International Journal of Teaching and Learning in Higher Education 2010, Volume 22, Number 3, 337-345
http://www.isetl.org/ijtlhe/ ISSN 1812-9129

Influence of the Pedagogical Context on Students' Evaluation of Teaching

Edna Luna, Vicente Arámburo, and Graciela Cordero
Universidad Autonoma de Baja California

The purpose of this study was to compare the characteristics of teaching performance in accordance with the opinion of students of different academic fields and curriculum stages in a Mexican state public university. The sample was composed of 729 randomly-selected courses, distributed over four semester periods. Descriptive and comparative statistical analyses were made. The results determined significant differences when natural-exact sciences were compared with administrative sciences ($p = .003$), and engineering with administrative sciences ($p = .022$) in the overall ratings and by dimension. Moreover, differences were found in the ratings by dimension between the curriculum stages. The study concludes in favor of considering the particularities of the pedagogical context in the interpretation of ratings, and of using them as a source of information when designing strategies for improving teacher training.

In the university, the evaluation of instruction plays a determining role in advancing the quality of learning. Key documents of Mexican and international educational policy recognize that instruction is important in achieving educational quality (Asociación Nacional de Universidades e Instituciones de Educación Superior, 2007; Anderson, 2004). The importance of evaluating instruction stems from its potential as a tool contributing to teachers' becoming professional and thus, improving their training.

Teacher evaluation based on Student Evaluations of Teaching (SETs) effectiveness is characterized by two particularities: it is the strategy most often used in North America, Europe and Asia, and is also the one most studied (Theall & Franklin, 2000; Seldin, 1993). In this context, two situations stand out regarding rating forms: first, there is a good deal of evidence for the misuse of the ratings students give; and second, teachers show a growing unease about the use of these ratings in the making of administrative decisions.

A great part of the criticism regarding rating instruction concerns the procedures of application, interpretation, and use of the results (Sproule, 2000; Díaz-Barriga, 2004). In particular, one of the most frequent errors related to the interpretation of the results is the aggregation of all the teachers' ratings without consideration for the particularities of the pedagogical context, such as the disciplinary field in which they teach and the educational stage.

The purpose of this work is to compare the characteristics of the teacher's performance, according to students' opinion, by disciplinary field (natural-exact sciences, engineering and technology, and administration sciences), and curriculum stage (basic and disciplinary/final stages). The objective is to contribute to the discussion regarding the interpretation and the use of the results of students' evaluations of university teaching.

In Mexico, as in other countries, the evaluation of teaching has resulted from social demands coming from different audiences with heterogeneous needs of evaluation and has been linked to the establishment of federal policies in this area. Until the end of the eighties, the evaluation of instruction was conducted primarily because of the institutions' need to obtain information on the quality of teaching and, in theory, to provide feedback on the strategies of teacher training (Arias, 1984; Luna, 2002). Since 1990, with the widespread implementation of merit pay programs, the evaluation of teaching has been included as one of the indicators of these programs. Hence, attention has been given principally to the need for administrative control over instruction (Canales & Gilio, 2008). Today, expectations for the evaluation of teaching are diverse: teachers and students expect fair and appropriate systems to improve teaching; the authorities seek to have better information for administrative decision making, allocation of courses, promotions and economic incentives; and governmental institutions seek a means of accountability for the quality of instruction (Luna, 2004; Secretaría de Educación Pública, 2007).

In Mexico however, research on the evaluation of university teaching is a recent development and is still in its infancy. The investigation into the evaluation of instruction began after evaluation policies were instituted at the beginning of the nineties, and it was in 1996 that the systematic production of literature regarding the topic began (Luna & Rueda, 2008). This is unlike the situation in other countries where there is a long history and tradition regarding SETs. Furthermore, the Mexican State has promoted an evaluation of teaching associated with policies of control and wage compensation, and as a result, this type of assessment has idiosyncrasies which have transcended research—for example, the difficulty of creating evaluation procedures apart from control.

In general, the main reasons for using the SETs are related to measuring the effectiveness of administrative decisions; the diagnosis and feedback of teachers to improve the process of instruction; and general research on teaching (Marsh and Dunkin, 1997). As a result, the ratings are considered useful for teachers, students, and administrators.

The practical and theoretical usefulness of the rating forms depends on complying with the psychometric standards for designing and applying the instrument. In the 80s and 90s, research was oriented toward studying the reliability and validity of the rating forms to measure teaching efficacy. Today, it has been demonstrated that the ratings of these instruments are reliable, stable, and relatively valid by means of their application in different educational scenarios (Abrami, D'Apollonia & Cohen, 1990; Marsh & Dunkin, 1997; Marsh, 2001).

One line of investigation in recent years studied the procedures of application, interpretation, and use of the results. This line of inquiry is important, as it has been shown that an incorrect procedure may invalidate the results (*v. gr*. Theall & Franklin, 1990). Therefore, we must emphasize the need to take great care regarding the validity and reliability of the instrument, as well as the credibility and fairness of the evaluation system.

As a result of research on validity, it is particularly relevant to investigate the impact of factors that affect the students' evaluation of the teacher, apart from the teacher him/herself. Although at the moment there is no consensus regarding a definition of bias in the ratings, an inclusive definition is that of Feldman (1997). This author defined bias in the ratings as one or more factors that directly and inappropriately influence the opinion of students about the evaluation of a course. The bias is determined based on the analysis of correlation between the opinion ratings and other variables. In classifying the factors that influence the students' evaluation of teachers, the following categories were identified: administration, characteristics of the course, characteristics of the instructor, characteristics of the students and characteristics of the instrument (Braskamp & Ory, 1994).

Research has also been done on the influencers of the results obtained from the rating forms for evaluation of instruction. The nature of the disciplinary field, the level of the course, and the size of the group in the classroom were found to have significant influence. Regarding the first, evidence obtained from the hierarchization of teachers' ratings has shown that students from different disciplinary fields evaluate in a differential manner. There exists a consensus that the ratings for teachers of english, humanities and the arts tend to be located in the upper and middle levels; for those in the social sciences (political science, sociology, economics and psychology) in the medium low; and the ratings for those of the natural-exact sciences, and engineering, in the low level (Cashin, 1990; Beran & Violato, 2005).

Differences in ratings between teachers in different disciplinary fields have been found both in the dimensions and in the overall evaluation. However, it is recognized that the results are not conclusive for determining the manner in which they differ (Hoyt & Lee, 2002). Furthermore, it should be emphasized that these studies are based upon particular SETs of universities in the United States of America, Australia, and Canada.

In Mexico, Luna & Valle (2001), and Luna (2002) studied the hierarchy of the dimensions of evaluation of instruction of teachers and students in the graduate programs of a public university. They found that trends of opinion do exist among teachers and students of the different programs, as regarding the preference of the dimensions, so that the groups obtained reflect a pattern of clusters by academic field in both populations. Garcia (2003) investigated the case of a private university and reported significant differences in teacher performance by academic field. The entire faculty of the Department of Humanities and Sciences of Mankind obtained averages that were higher for teachers of the science and engineering department, economic and administrative sciences, and the arts.

Regarding the influence of the educational level on the ratings, early works investigated the overall effectiveness of the instructor in relation to the level of the course—the first semester compared with the last, and found no differences (Erdle & Murray, 1986). However, research carried out in the 90s analyzed the ratings given to the dimensions and concluded that differences do exist between the ratings. For example, Smith & Cranton (1992) found that for students in the early years of college, the organization of the course and the clarity of exposition are the most important dimensions. For advanced and graduate students, the most important aspects are the atmosphere in the classroom and the evaluation of learning. Hativa (1996) found that first semester students place more importance on aspects related to the teacher's interaction with students, while students in advanced semesters attach greater significance to the teacher's mastery of the subject s/he teaches.

One common limitation of SETs is that they concentrate teachers' ratings in one group without recognizing the individual characteristics of the teaching context (Stake & Cisneros, 2000). For example, it has been demonstrated that it is possible to compare the ratings obtained by teachers from different disciplinary fields, with groups of the same size and level of education, only when it is clear that the ratings between the schools or departments are similar and there are no important differences between the means of

Table 1
Distribution of the Courses that Make Up the Sample
by Academic Field and Curriculum Stage

Nature of the Discipline	Academic Field	Curriculum stage		Total Courses
		Basic	Disciplinary/Final	
Hard Pure Sciences	Natural-exact	52	70	122
Hard Applied Sciences	Engineering and Technology	141	114	255
Soft Applied Sciences	Administrative	174	178	352
Overall Total		367	362	729

the questions (Theall & Franklin, 2000). In Mexico, no studies have been conducted to analyze the impact of teaching context on teacher performance, nor are there any concrete guidelines as to how these results should be interpreted to improve teaching practice.

The Context of the Study

This study was performed at the Autonomous University of Baja California (UABC) in Mexico. The UABC offers 57 undergraduate degree programs, and 48 postgraduate, distributed over various disciplinary fields. There are 36,432 undergraduate students enrolled, and there are 1,516 full-time faculty members (Universidad Autónoma de Baja California, 2007). At the UABC, curricula are organized by stages of training. The stages are basic and disciplinary/final. The basic stage comprises the first three semesters, while the disciplinary/final stage is the fourth through the eighth or ninth semester.

The evaluation of teaching, based on rating form answers provided by students, began in a systematic manner in 1988, with the purpose of obtaining information to be used in reorienting the training and development of the academic staff. In 1994, the ratings the students provided the teachers were added as one of the components of the academic staff's economic stimulus program (*merit pay*), which transformed the use of the results from what had been originally planned.

Since that year, reports have been provided to each academic unit so that the directors and the teachers themselves can use evaluation results. Moreover, the results are compiled in a database in a central administration office of the university as part of each teacher's record for the economic stimulus program. In other words, the institution's scores are primarily used for purposes of administrative control.

A central university office manages the SET process. Students answer the computerized rating form at the end of each semester period. A sense of scope regarding the amount of data generated by this process can be gained by considering the first semester of 2008. A total of 28,210 students (74.6% of the enrollment) assessed 3,629 teachers during that time period (S. Osuna, personal communication, October 20, 2008).

The objectives of this study were:

1. To compare the characteristics of teaching performance according to the academic field: natural-exact sciences, engineering and technology, and administrative sciences.
2. To compare the characteristics of the ratings, according to the stage of training (basic, and disciplinary/final).

Therefore, the central questions for study were: 1) Are there differences in the ratings students give to teachers, according to the disciplinary fields to which they belong? 2) Are there differences in the ratings assigned by students to teachers according to the students' curriculum stage?

Method

Source of Data and Sample

The data used in this study were obtained from UABC undergraduate courses delivered during the four semester periods of 2004 and 2005. Specific disciplinary areas were considered. They were natural-exact sciences (BS in Physics, BS in Mathematics and BS in Biology), engineering and technology (Civil Engineering, Electronic Engineering, Industrial Engineering, Computer Engineering), and administrative sciences (BA in Accounting, BS in Computer Science and BA in Business Administration). These disciplinary fields were chosen as criterion of comparison because of an interest in contrasting the results of teacher evaluations between pure hard sciences (natural-exact sciences), hard applied sciences (engineering and technology) and soft applied sciences (administrative sciences) (Biglan, 1973).

The selection of the sample was made according to the following criteria: a) included were all courses which have been evaluated by a minimum number of students, based on the reliability indices proposed by Centra (1993); b) of the total courses fulfilling the

above criteria in the areas of engineering and administration, 30% were selected at random; c) in natural-exact sciences, the criterion was to have a minimum of 30 courses, since there were few records that met the requirement for inclusion. The distribution of the sample by academic field and training stage is shown in Table 1. The overall sample was composed of 729 courses.

This investigation employed a retrospective and comparative study methodology (Mendez, Namihira, Moreno & Sosa 2001). It is retrospective because it analyzed evaluation ratings for courses given during periods prior to the study. It is comparative because comparisons were made to identify characteristics of teacher performance, according to students' opinion by academic field and curriculum stage.

Instrument and Variables

The ratings were collected using the *Rating Form for the Evaluation of Teaching* designed *ex profeso* for the UABC. The instrument contains 20 questions: 2 closed-response and 18 Likert-type. It focuses on eight dimensions of teaching: 1) structure of objectives and content; 2) clarity of instruction; 3) organization of the class; 4) mastery of the subject; 5) teaching strategies; 6) quality of interaction; 7) evaluation of learning; and 8) work methods.

The rating form also includes information that allows the identification of the course and the teacher, such as: course name, degree program to which it belongs, and teacher's name. According to the study done on the psychometric characteristics of the instrument, it was concluded that it belongs to the theory of cognitive learning, and has a reliability index of 0.94 and a 75% percentage of explained variance (Luna & Valle, 2005).

Students answer the rating form at the end of each semester period using a computer. The ratings are concentrated in a database of the university (central administration) that processes the ratings and provides reports for each subject. These include average ratings by dimension and the overall average of the course on a scale of 1 to 10. In addition, these reports identify the course, the major to which it belongs, and the teacher evaluated.

In this study, the variables considered were academic field of the course; curriculum stage of the course; overall average of the course ratings; and average score for the dimensions.

Procedure

The procedure was developed in two phases:

Phase 1: Design of the software for processing the information. Because the students' ratings are concentrated in a university database, a program was designed to allow us to obtain the information from that database and to organize it according to the variables of interest. The program is linked to the institutional database and collects the necessary information. Information required for this study, but not contained in the university database—specifically the identification phase of curriculum courses—was fed in manually.

Phase 2: Statistical analysis. Calculations were made with the Statistical Package for the Social Sciences (SPSS), and the analyses were descriptive and comparative. The first type consisted in analyzing the variations in teaching performance by academic field, training stage in each of the scholastic periods, and the sum of the four periods. To do this, the arithmetic means and standard deviations of the ratings were calculated by academic field and training stage. The average ratings for each of the eight dimensions of teaching by academic field and curriculum stage were also calculated.

The first comparative analysis considered the four scholastic periods by academic field, and Analysis of Variance (ANOVA) was used. Afterward, the following tests were used: F-Levene (to detect homogeneity of variance); one-way ANOVA (to compare differences between the mean ratings of the three academic areas and between dimensions); the *post-hoc* analysis (Tamhane, DunnetT3 and Tukey - to locate significant differences); and finally, a *t-student test* (to contrast the means of the overall ratings of the courses and the dimensions for the basic and disciplinary/final stages).

Results

The results are presented in two blocks, according to the objectives of the research, by academic area and by curriculum stage. The dependent variables are: overall average of the course ratings and average of the ratings of the teaching dimensions. The independent variables are: academic field and the curriculum stage of the course.

Results by Academic Field

ANOVA was used to compare the average ratings of the four scholastic periods. No significant differences were found between the four periods. This suggests that there is stability in the ratings given to the teachers over time. In Table 2, we see the average ratings by disciplinary fields. This demonstrates the significant differences revealed by the ANOVA test.

Table 2
Comparisons of Overall Averages by Academic Area,
and Values and Levels of Significance in the ANOVA Tests

Disciplinary Fields	Basic Descriptives		Homogeneity of Variance	ANOVA	
	$\overline{X}$	*s*	Significance	*F*	*Sig.*
Natural Sciences	8.91	0.68	0.011	6.64	0.00
Engineering	8.98	0.85			1
Administrative Sciences	9.15	0.64			

Table 3
Differences Between Academic Areas,
Shown by the Execution of the Post Hoc Analysis

Comparisons by Areas of Knowledge	Levels of Significance
Natural-exact *vs.* engineering	0.804
Natural-exact *vs.* administrative	0.003*
Engineering *vs.* natural-exact	0.804
Engineering *vs.* administrative	0.022*
Administrative *vs.* natural-exact	0.003*
Administrative *vs.* engineering	0.022*

* Values with statistical significance of level $p < 0.05$.

Figure 1
Averages for Teacher Performance by Dimension and Academic Field

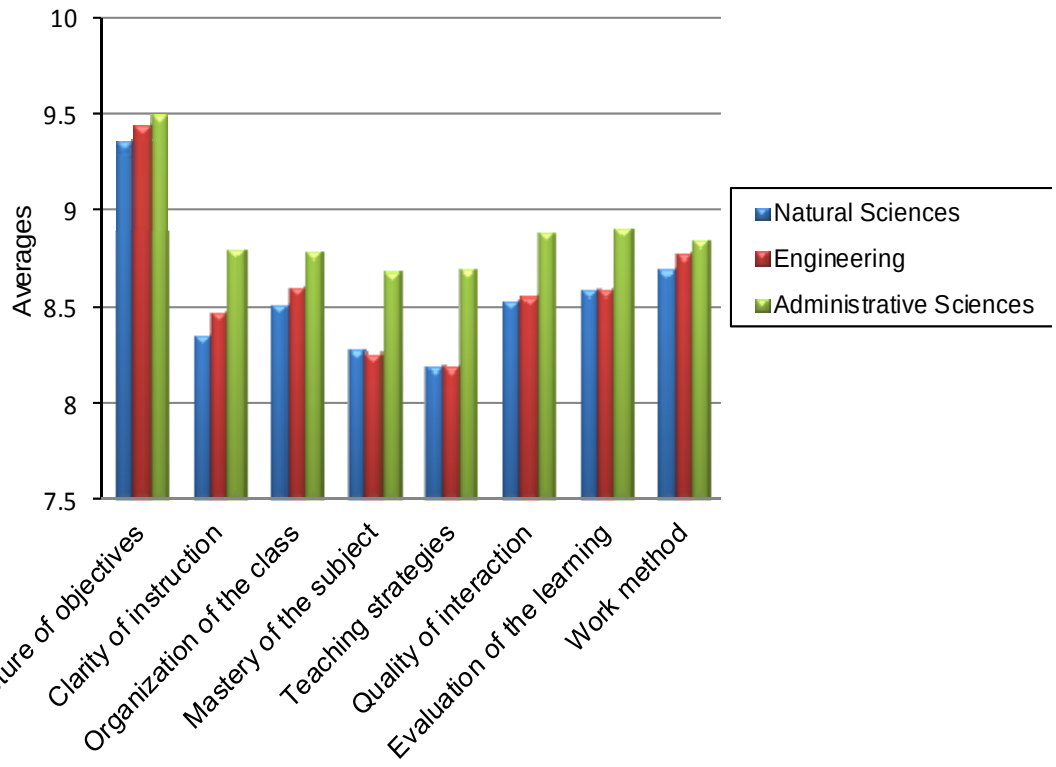

Table 4
Comparisons Between Basic and Disciplinary Stages, by Dimension, for the Three Areas of Knowledge, by Means of the t-Student Test

Academic Area	Dimension	Curriculum Stage	*n*	$\bar{X}$	*t*	*Sig.*
Natural-exact Sciences	Organization of the Class	Basic	52	8.28	-2.52	0.013*
		Disciplinary/final	70	8.69		
	Mastery of the Subject	Basic	52	8.0	-2.48	0.014*
		Disciplinary/final	70	8.48		
	Teaching Strategy	Basic	52	7.88	-2.79	0.006*
		Disciplinary/final	70	8.43		
	Quality of Interaction	Basic	52	8.24	-2.42	0.017*
		Disciplinary/final	70	8.70		
Administrative Sciences	Quality of Interaction	Basic	174	9.0	3.13	0.033*
		Disciplinary/final	178	8.79		

When the *post hoc* analysis was executed, statistical differences were found between natural-exact sciences and administrative sciences, and between the areas of engineering and administrative sciences (see Table 3). In contrast, no differences were presented between the areas of natural-exact sciences, and engineering. The results of the foregoing analysis affirm that students in the field of administrative sciences evaluate the teacher with significantly higher ratings, followed by those of engineering and the natural-exact sciences.

The averages of the ratings of the dimensions of teaching and the academic areas are presented in Figure 1. In the three areas, *structuring of objectives and content* is conspicuous as the best-evaluated dimension by students. By contrast, the lowest averages coincide in the three disciplinary fields in the dimensions *mastery of the course* and *teaching strategies.* These are fundamental teaching functions.

The results of the ANOVA indicated significant differences in six of the dimensions of teaching by academic field. The differences according to the post hoc were between administrative sciences and the other two areas, particularly in: *clarity of instruction* (F = 14.4, p=.000), *organization of the class* (F = 5.03, p = .007), *mastery of the subject* (F = 9.2, p = .000), *teaching strategies* (F=15.1, p=.000), *quality of interaction* (F = 12.1, p= .000), and *evaluation of learning* (F = 8.9, p = .000). Thus, there were no significant differences in the dimensions *structure of objectives* and *work method.* Hence, it is confirmed that administrative science is the area best evaluated, both in the overall averages and in averages by dimension.

Results by Curriculum Stage

Comparative analysis of the overall analyses between these two stages resulted in no significant differences. Similarly, comparison of the overall averages by training stage and academic area did not report significant differences.

The results of the t-student test of the dimensions by curriculum stage revealed significant differences in four dimensions of natural-exact sciences and in one of the administrative sciences (see Table 4). In natural-exact sciences, students of the disciplinary/final stage gave better ratings to teachers than did students of the basic stage, particularly in the dimensions *organization of the class, mastery of the subject, teaching strategies,* and *quality of interaction.* In administrative science, only in the dimension *quality of interaction* were significant differences found. Finally, in engineering and technology, no significant differences were found in any of the dimensions of any of the stages contrasted.

Discussion

Since the nineties, the literature has insisted upon the use of SETs as part of a broader system of teacher evaluation. Implicit in this is a number of basic requirements, including an explicit articulation of the purposes of evaluation; assurance, in the administrative management, of the reliability of the process; and a determination of those actions that would lead to an improvement in teaching practices. Moreover, there is an emphasis on the need for linking the findings of SETs with evaluation systems for the purpose of improving evaluation practices.

Recognition of the complexity of the evaluation of instruction obliges us to investigate the various elements involved in the teaching/learning process, as well as to determine the importance of the elements of interaction and its principal effects. The context in which evaluation takes place has shown that it has serious effect, both in interpretation and in the usefulness of the students' ratings. In this study, we hoped to make advances in the analysis of the results of the students' evaluations of the teacher, according to the academic field and the curriculum stage of their training.

In this work, stability was found in the ratings for the four periods studied; this is consistent with previous

research (Abramian, D'Apollonia & Cohen, 1990; Marsh & Dunkin, 1997; Marsh, 2001). Traditionally, this datum has been used only to justify the reliability of the ratings. However, in the framework of a system of teacher evaluation, it should be considered as empirical information for finding out the individual strengths and weaknesses of teachers and groups of teachers, by disciplinary fields.

One of the findings of this study concerned the differences in the ratings given by students to teachers according to their academic field. It was discovered that teachers of the pure hard sciences (natural-exact) as well as those of hard applied sciences (engineering and technology) received lower ratings than teachers of soft applied science (administration). These findings concur with those observed in other educational environments (Hativa, 1996; Beran & Volato, 2005) and must be considered when analyzing and interpreting the ratings for administrative purposes, if we expect the system to be fair to teachers.

Concerning students' perceptions of teacher performance as related to the curriculum stage, significant differences were found only in natural-exact sciences in analysis by dimension. Disciplinary/final stage courses received higher ratings. Other studies have found that the courses of the last semesters of undergraduate and postgraduate programs tend to receive higher ratings (Marsh & Hocevar, 1991; Marsh & Dunkin, 1997). In this regard, Marsh & Dunkin (1997) argue that the effects of the stage of the course tend to disappear when other prior variables are controlled, although these findings are difficult to interpret, given that there a specific model does not exist for organizing the variables.

In this work it is assumed that evaluation procedures must be sensitive to the complexity of teaching. It is likewise assumed that teaching can be judged in an appropriate manner only if it is evaluated within the framework of the factors that determine it. From this point of view, it is expected that the systems be differentiated, and they be congruent with their educational context and with the characteristics of the teachers. The results of this work support this expectation, since they consistently show differences in the ratings students give teachers in the various disciplinary fields.

The particularities of the teaching process in different disciplinary settings; as well as the particularities of the context, such as the type of course, the size of the group and the characteristics of the teacher, must be investigated in future studies for the purpose of gaining a better understanding of the factors which affect teacher competence. Various authors have noted the maneuvering power wielded by evaluation processes, and specifically, it has been argued that an adequate program of teacher assessment using rating forms can lead to the improvement of teaching (Marsh, 2001). Although the ratings obtained are useful for teachers, students, administrators, and for improving educational practices, the possibilities for the extrapolation of the results depend on the way technical factors (described extensively in this article) and organizational factors (Darling-Hammond, 1997) interact.

In terms of organizational factors, it is important to mention that the rating form results cannot be applied for the improvement of teaching unless certain things are taken into consideration. According to Centra (1993), the processes of teacher evaluation can support the improvement of teaching if they meet four critera: they must provide the teacher with new information, permit the teacher to value the information, provide the teacher with strategies for improving his/her performance, and motivate the teacher to make improvements. Similarly, for Seldin (1993) the usefulness of rating forms depends on two factors: that the teachers feel personally motivated to improve, and they know how to improve. The university, as an organization, must integrate these considerations into its improvement endeavors.

Another crucial aspect of organization is the way in which this information is made known to the teacher. At an empirical level, changes have been demonstrated in the effectiveness of teaching based on feedback derived from rating forms; however, it is important to note that modifications are minimal when the results of the ratings are only provided in writing. Feedback can have a far great impact when it is accompanied by a personal interview (L'Hommediu, Menges & Brink, 1990). The data provided by the ratings given by students provide an information base for delimiting skills to develop in the teacher-training programs, insofar as they report the strengths and weaknesses of the teaching task.

Undoubtedly, the information derived from student ratings should be part of an overall diagnosis of needs, which furthermore can be complemented with other methodologies which would explore particular needs in detail (Luna, Cordero & Galaz, 2007). In this sense, the authors concur with Duke & Stiggins (1997) in the sense that this type of proposal should nurture plans for evaluation that would have as their fundamental objective the professional development of our teachers.

References

Anderson, L. (2004). *Increasing teacher effectiveness*. Paris, FR: UNESCO: International Institute for Educational Planning.

Abrami, P. C., D'Apollonia, S., & Cohen, P. A. (1990). Validity of student ratings of instruction: What we know and what we do not. *Journal of Educational Psychology, 82*(2), 219-231.

Arias, G. F. (1984). El inventario de comportamientos docentes (ICD): Un instrumento para evaluar la calidad de la enseñanza. *Perfiles Educativos, 4,* 14-22.

Asociaciòn Nacional de Universidades e Instituciones de Educación Superior. (2007). Consolidación y avance de la educación superior en México. MX: Author.

Beran, T., & Violato, C. (2005). Ratings of university teacher instruction: How much do student and course characteristics really matter? *Assessment and Evaluation in Higher Education*, *30*(6), 593–601.

Biglan, A. (1973). The characteristics of subject matter in different academic areas. *Journal of Applied Psychology*, *57*(3), 195-203.

Braskamp, L. A., & Ory, J. C. (1994). *Assessing faculty work: Enhancing individual and institutional performance.* San Francisco, CA: Jossey-Bass.

Canales, A., & Gilio, M. C. (2008). La actividad docente en el nivel superior ¿diferir el desafío? In M. Rueda (Coord.), *La evaluación de los profesores como recurso para mejorar su practica* (pp. 17-38). MX: Universidad Nacional Autónoma de México, IISUE.

Cashin, W. E. (1990). Students do rate different academic fields differently. In M. Theall & J. Franklin (Eds.), *Student rating of instruction: Issues for improving practice* (pp. 113-123). San Francisco, CA: Jossey Bass.

Centra, J. A. (1993). *Reflective faculty evaluation.* San Francisco, CA: Jossey-Bass.

Darling-Hammond, L. (1997). Evolución en la evaluación de profesores: nuevos papeles y métodos. In J. Millman & L. Darling-Hammond (Eds.), *Manual para la evaluación del profesorado* (pp. 23-45). Madrid, ES: La Muralla.

Díaz-Arceo, F. (2004). Algunas críticas en torno a los métodos de evaluación de profesores y algunas Incursiones alternativas. In M. Rueda & F. Díaz-Barriga (Coords.), *La evaluación de la docencia en la universidad* (pp. 121-134). MX: Universidad Nacional Autónoma de México-Plaza y Valdés.

Duke, D., & Stiggins, R. (1997). Más allá de la competencia mínima: Evaluación para el desarrollo profesional. In J. Millman & L. Darling-Hammond (Eds.), *Manual para la evaluación del profesorado* (pp. 165-187). Madrid, ES: La Muralla.

Erdle, S., & Murray, H. (1986). Interfaculty differences in classroom teaching behaviors and their relationship to student instructional ratings. *Research in Higher Education*, *24*, 115-127.

Feldman, K. A. (1997). Identifying exemplary teachers and teaching: Evidence from student ratings. In R. Perry & J. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 368-395). New York, NY: Agathon Press.

García, G. J. (2003). Profesores universitarios y su efectividad docente. Un estudio comparativo entre México y Estados Unidos. *Perfiles Educativos*, *25*(100), 42-55.

Hativa, N. (1996). University instructors' ratings profiles: Stability over time, and disciplinary differences. *Research in Higher Education*, *37*, 341-365.

Hoyt, D., & Lee, E. J. (2002). *Disciplinary differences in student ratings* (Technical report No. 13). Manhattan, KS: The Individual Development and Educational Assessment Center. Retrieved from http://www.theideacenter.org/sites/default/files/techreport-13_0-pdf.

L' Hommedieu, R., Menges, R. J., & Brinko, K. T. (1990). Methodological explanations for the modest effects of feedback. *Journal of Educational Psychology*, *82*, 232-241.

Luna, E. (2002). *La participación de los docentes y estudiantes en la evaluación de la docencia*. MX: Universidad Autonoma de Baja California-Plaza y Valdés.

Luna, E. (2004). Los cuestionarios de evaluación de la docencia por parte de los alumnos recomendaciones para su utilización. In M. Rueda & F. Díaz-Barriga (Coords.), *La evaluación de la docencia en la universidad* (pp. 98-121). MX: Universidad Nacional Autónoma de México-Plaza y Valdés.

Luna, E., & Valle, C. (2001). Diferencias y similitudes en las opiniones de docentes y estudiantes sobre las dimensiones de la enseñanza efectiva. In M. Rueda Beltrán, F. Díaz Barriga, & M. Díaz Pontones (Eds.), *Evaluar para comprender y mejorar la docencia en la educación superior* (pp. 113-123). MX: Universidad Autónoma Metropolitana-Universidad Nacional Autónoma de México-Universidad Autónoma Benito Juárez de Oaxaca.

Luna, E., & Valle, C. (2005). Características pedagógicas y validación de un instrumento para la evaluación de la docencia universitaria. In *Proceedings of the VIII Congreso Nacional de Investigación Educativa* [CD- ROM]. MX: Consejo Mexicano de Investigación Educativa-Universidad de Sonora.

Luna, E., Cordero, G., & Galaz, F. (2007). Los resultados de la evaluación docente y su uso para el diseño de modalidades de formación de los profesores. Paper presented at *I Congreso Internacional Nuevas Tendencias en la Formación Permanente del Profesorado*, Barcelona, Spain.

Luna, E., & Rueda, M. (2008). Estado del conocimiento sobre la evaluación de la docencia universitaria

1990-2004. In M. Rueda (Coord.), *La evaluación de los profesores como recurso para mejorar su practica* (pp. 39-58). MX: Universidad Nacional Autónoma de México-IISUE.

Marsh, H. W. (2001). *Students' evaluations of university teaching*. Retrieved from http://apps.uws.edu.au/uws/edc/seeq/SETs_HerbMarsh_presentation_2001.pdf.

Marsh, H. W., & Dunkin, M. J. (1997). Students' evaluations of university teaching: A multidimensional perspective. In R. Perry & J. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 241-320). New York, NY: Agathon Press.

Marsh, H. W., & Hocevar, D. (1991). Students' evaluations of teaching effectiveness: The stability of mean ratings of the teachers over a 13-year period. *Teaching and Teacher Education, 4*, 9-18.

Méndez, I., Namihira, D., L., & Sosa, J. (2001). *El protocolo de investigación. Lineamientos para su elaboración y Análisis*. MX: Trillas.

Seldin, P. (1993). The use and abuse of student ratings of professors. *The Chronicle of Higher Education, 39*, 41.

Secretaría de Educación Pública. (2007). *Programa Sectorial de educación 2007-2012*. (pp. 9-23). MX: Author.

Smith. R. A., & Cranton, P. A. (1992). Students' perceptions of teaching skills and overall effectiveness across instructional settings. *Research in Higher Educación*, *33*, 747-764.

Sproule, R. (2000). Student evaluation of teaching: A methodological critique of conventional practices. *Education Policy Analysis Archives*, *8*(50), 1-34.

Stake, R. E., & Cisneros, E. J. (2000). Situational evaluation of teaching on campus. *New Directions for Teaching and Learning, 83*, 51-72.

Theall, M., & Franklin, J. (1990). Student ratings in the context of complex evaluation systems. In M. Theall & J. Franklin (Eds.), *Student rating of instruction: Issues for improving practice* (pp. 17-34). San Francisco, CA: Jossey Bass.

Theall, M., & Franklin, J. (2000). Creating responsive student ratings systems to improve evaluation practice. *New Directions for Teaching and Learning, 83*, 45-105.

Universidad Autónoma de Baja California. (2007). *Informe de Rectoría 2007*. Retrieved from http://www.uabc.mx/planeation/informe/informe2007/informe2007.pdf.

EDNA LUNA, Ph. D., Institute of Educational Research and Development of the Baja California Autonomous University (Mexico); full-time professor and coordinator of the Doctor´s Degree Program in Educational Sciences; member of the Mexican National System of Research since 2000; member of the National Conuncil of Research Education.

VICENTE ARÁMBURO, M. Sc., Departament of Administrative and Social Sciences of the Baja California Autonomous University (Mexico), full-time teacher.

GRACIELA CORDERO, Ph. D., Institute of Educational Research and Development of the Baja California Autonomous University (Mexico); full-time full-time professor; currently dean of the Institute of Educational Research and Development; member of the Mexican National System of Research since 1999; consultant of international projects supported by the Canadian of International Development Agency (CIDA), and by the Agencia Española de Cooperación Internacional para el Desarrollo (AECID).

Acknowledgements

English translation by Lessie Evona York-Weatherman. This work was supported by the National Council of Science and Technology (of Mexico), project No. 49052.