

MODULATING FACTORS IN THE PERCEPTION OF TEACHING

[Factores moduladores de la percepción de la calidad docente]

by/por

[Article record](#)

[About authors](#)

[HTML format](#)

Casero, Antonio (a.casero@uib.es)

[Ficha del artículo](#)

[Sobre los autores](#)

[Formato HTML](#)

Abstract

This article examines the existing research findings in relation to the research of the factors that are presumably involved in university students' evaluations of the teaching quality they receive, possibly threatening the validity of the construct. The review is organized around the three implicated sources in the perceptive process of the student, being the student himself, the professor, and the context. The review concludes that, despite some minor effects, the factors under analysis do not represent a substantial threat to the system's validity. The unproductive debate that has gone on for decades about guarantees in the evaluation of teaching staff has largely been used to cover up the absence of a desire to institutionalize this kind of evaluation system.

Keywords

Student evaluation of Teacher Performance, Teaching Quality, Effective Teaching.

Resumen

El presente artículo tiene como objetivo presentar una revisión sobre los hallazgos existentes en relación a la investigación de los factores supuestamente implicados en la valoración que el alumnado universitario realiza sobre la tarea docente de sus profesores, pudiendo suponer una amenaza a la validez de dicho constructo. La revisión se presenta organizada en torno a las tres fuentes implicadas en el proceso perceptivo del alumno, siendo éstas el propio alumno, el profesor y el contexto. La revisión concluye que, a pesar de efectos menores, los factores analizados no representan una amenaza substancial a la validez del sistema. El debate infructuoso que se ha producido durante décadas sobre las garantías en la valoración del personal docente ha sido usado en gran parte para encubrir la falta de deseo de institucionalizar este tipo de sistemas de evaluación.

Descriptores

Evaluación del desempeño docente, Calidad docente, Efectividad docente

The opinion survey applied to university students to evaluate the effectiveness of their instructors has been used for eighty years now, but is still being challenged by substantial sectors of the teaching faculty, those with strong views on certain factors that would bias or unduly influence students' opinions of their instructors' teaching. This area - attacks on the validity of teaching evaluation

questionnaires - has been copious in terms of scientific output and concluded that none of the alleged bias factors show a considerable effect, although the odd weak impact has been detected (Apodaca and Rodríguez, 1999).

This study presents a review of the research studying these suspect variables,

which we dare not call bias as to deserve the description of “source of error” they must have an external influence substantially related to the evaluations and relatively unrelated to other teaching quality indicators. Or as Villa indicated, “To be able to prove that a characteristic nullifies the evaluation, a significant correlation is not enough even if a causal relationship is plausible; this feature must also not be positively related to other indicators of effective teaching.” (Villa, 1985, 84).

The basic elements of the relationship studied here are the student and the instructor. The latter performs an action that is evaluated in one way or another by the former. And we cannot ignore the setting, which is a third element that modulates the relationship. Thus, we understand that student judgments are the product of their perceptual process, stimulated by the behaviour of the perceived object: the instructor. This process takes place in a setting with its own spatial and temporal characteristics. Thus, these three elements make up the sources of variation in the perceptual process, as posited by social psychology in its studies on human perception.

The scheme that will be used in this presentation consists in assigning variables to each of the three sources involved in students’ perceptual process, these sources being the students themselves, the instructor and the setting.

Student variables

The following section covers the student variables that have been studied as possible modulators of responses in research on teaching evaluation questionnaires.

Gender

Aparicio, Tejedor and Sanmartin (1982) cited a number of studies from the US that yielded contradictory results. Some authors found no differences between male and female students; others indicate that female

students are more critical than males and that they tend to rate their female instructors higher; still others assert that female students give higher rating to their instructors on some items than males students.

Aleamoni and Hexner (1980) cited similar references and Marsh (1987) reached similar conclusions.

Fernández and Mateo (1997) carried out a variance analysis with three factors: student gender, instructor gender and faculty (the faculty of sciences, faculty of social sciences or technical schools). “Teaching competence” and “motivation/interaction skills” were the two factors extracted by an exploratory and confirmatory factor analysis of the abridged version of the Complutense University’s Instructor Evaluation Questionnaire (CUTEQ-R) used as dependent variables in the 2 x 2 x 3 design. The ANOVA results showed three statistically significant components: student gender, which had an effect on both teaching competence and motivation and interaction skills, with female students rating instructors higher; and the faculty’s effect on the motivation and interaction skills - higher evaluations at the faculty of social sciences, followed by the technical school and, lastly, the faculty of sciences. The size of the effects was truly low; 0.0078, 0.0098 and 0.0072, respectively, were the omega-squared indices found. To sum up, as the authors noted, the effect of student gender on evaluations of the quality of their instructors is very weak or almost nonexistent.

Expectations

Although the effects of instructor expectations on students are better known, student expectations can also influence their own behaviour and the instructor’s. Feldman and Prohaska (1979) conducted two experiments:

In the first, an instructor gave a class to a group of university students. Half of them were told that the instructor was very competent and the other half were told the opposite. Significant differences in the attitudes, per-

formance and nonverbal behaviour of the groups were found. The students who expected a competent instructor described the lesson as easy and interesting; they evaluated the instructor as more competent, intelligent and enthusiastic; they scored higher on a performance test and displayed different verbal behaviour, characterised by leaning in more toward the instructor and looking at him/her more often.

In the second experiment, a group of university students were asked to demonstrate the nonverbal behaviour described above – tilting the head and looking at the instructor – and another group was asked to demonstrate the opposite behaviour in a class with an instructor, who was actually another student pretending to be an instructor. The “instructors” who taught class to students with positive nonverbal behaviour felt better during it, perceived themselves as more competent, viewed the students as friendlier and more enthusiastic, were more effective, and showed a different nonverbal behaviour characterised by leaning in less toward the students than the other subjects did.

Teacher Folklore

Closely linked to the idea of expectations are a number of studies that examine the “teacher folklore” formula. These studies analyse the effect of the student’s opinion at the start of the course and before meeting the instructor on the end-of-course evaluations. The pre-existing idea of the instructor was communicated by students in the highest level courses. Reputational aspects, sometimes referred as “entry attitudes”, are also usually included in this variable.

Aleamoni et al. (1972) studied the teacher folklore operative in course content, workload, teaching style, etc., and concluded through pretest-posttest correlations that it does not contribute significantly to course ratings.

Another study conducted by Barké, Tollefson and Tracy (1983) attempted to link entry

attitudes and end-of-course evaluations. To do so, they selected 75 instructors at random stratified by age and rank from a total of 500. These instructors were evaluated by 789 student respondents to a pre-test and post-test. The results indicated that only a minority of students expressed expectations about the instructor and course, and that instructor ratings were moderately predictable in these few cases. Thus, “instructors can be victims - or beneficiaries - of their reputations” (Barké, Tollefson and Tracy, 1983, 83). According to Marsh (1984) expectations can affect evaluations, but for most students, the evaluations reflect their judgements of the quality of teaching.

Prior subject interest

The studies available in the area of prior subject interest have been conducted in relation to the Students’ Evaluations of Educational Quality (SEEQ) instrument by Marsh and others who, in turn, reviewed earlier studies. Prior subject interest correlated with the learning/value dimension, $r = 0.40$. “Student interest facilitates effective teaching and creates a more favourable learning environment and this effect is reflected in student evaluations.” (Marsh, 1987, 315).

The results do not clarify whether the same instructor teaching a subject without the students’ prior interest would obtain the same ratings.

Student personality

A study by Abrami and d'Apollonia (1990) presents three studies that examine the relationships between students’ personality traits, instructor evaluations and student performance. In the first study, 388 students evaluated themselves on the Adjective Check List (ACL) (Gough, 1979) and then watched one of four videos that varied in instructor expressiveness (low, high) and content of the explanation (low, high). Afterwards, they evaluated the instructor in the video and took a test based on the explanation. Lastly, they completed the ACL on the instructor. The

second study was similar, except that the 87 students in this case watched two videos, each a week apart. In the third study, 108 students from five Introduction to Psychology classes completed the ACL about themselves, evaluated their instructor's teaching, took a test on the common course material and completed the ACL on their instructor. The findings suggest four conclusions relevant to the summative evaluation of teaching (Abrami and d'Apollonia, 1990, 111):

- a) There do not appear to be any significant or consistent relationships between students' personalities and ratings.
- b) The instructors' personality traits as perceived by students are related to ratings of teaching effectiveness.
- c) Evaluations better predict the instructor's performance for classes in which the personality traits of the enrolled students differ.
- d) The instructor's effects on ratings were significantly higher than these effects on performance.

Evaluation styles

An interesting study by Castro (1996) analyses the existence of what he calls "evaluation styles" to refer to stable student evaluation patterns, i.e., a permanent evaluation style that always, or almost always, views reality evaluated in the same way, regardless of what is being evaluated. The author based his hypotheses on cognitive styles and the theory of cognitive-affective molds (Hernández, 1991).

The study was conducted with 6,040 students at the University of Las Palmas de Gran Canaria, to whom four instruments were administered:

- An instructor evaluation questionnaire: 35 items, yielding 6 factors that explain 53% of the variance. A total of 23,816 evaluations were conducted, in which each student evaluated four instructors on average.

- A background elements evaluation questionnaire: composed of 14 items that evaluate the degree of student satisfaction with a number of elements that make up their immediate academic context: relationship between students, ratios, university services, etc. The instrument yielded four factors that accounted for 47.7% of the variance.

- A causal attribution questionnaire on the poor quality of university teaching: comprising 14 items related to possible causes of poor teaching quality. The instrument yielded four factors that explained 53.1% of the variance.

- A causal attribution of student academic failure questionnaire: made up of 11 items related to the possible causes of academic failure. The instrument yielded three factors that explain 48.3% of variance.

The study's findings reveal, among other things, "the students' efficacy to be unbiased, reliable, valid and objective evaluators for the most part, and ample capacity for discriminating among different instructors and between these and other situations at the same time has been demonstrated" and that "there is a student evaluation style with respect to the faculty that allows a minority of students to be classified as:

A. *Critical*, those who rate their instructors lower, are more dissatisfied with their academic context and attribute more responsibility for academic failure and the poor quality of teaching to elements most closely associated with the faculty;

B. *Benevolent*, those who behave in the opposite way from the preceding point; or

C. *Even-handed*, the majority of students, whose ratings fall between the two previous groups." (Castro, 1996, 471)

The author concludes by recommending the use of control items to predict the class' general evaluation style in order to ascertain whether an extreme style dominates or, conversely, if all styles are equal.

Subsequently, Etopa (2003) analysed the results of surveys at the University of Las Palmas in Gran Canaria and compared them using two indicators of central trends: the arithmetic mean and the trimmed mean at 5%. The latter is more robust than the former in eliminating the 10% end from the calculation. The results led him to conclude that “eliminating the evaluations in which a benevolently-critically biased judgement is issued does not significantly increase the mean and reinforces that ULPGC students are valid and reliable evaluators” (Etopa, 2003, 269).

Expected and/or received grades

Peterson and Cooper (1980) indicated that several studies have found a positive correlation between the grade or mark a student expects or receives, and evaluations of the instructor they are rating. A belief exists that students reward instructors who give them good grades and punish those who do not. The authors conducted a study with two groups of instructors: one that graded their students and another that simply did or did not give them credits; the authors compared the evaluations of the two groups and concluded that they were similar.

For Marsh (1980), the expected grade is the background variable that may reasonably involve a bias in the ratings, although this interpretation is subject to alternative explanations. On the students' part, previous subject interest may be the basis for a better learning experience conducive to better grades; on the instructor's part, more effective teaching may be leading to better student learning and higher grades. Furthermore, since workload positively correlates with ratings, it may be that instructors who give students higher workloads are rated higher and students, in turn, perform better and receive higher grades.

Aleamoni and Hexner (1980) cite many studies that illustrate the controversy surrounding student rating and expected or re-

ceived grades. The general trend is that students tend to rate courses and instructors more positively when they receive or expect to receive good grades. They offer a median correlation of approximately 0.14, with a mean and deviation of 0.18 and 0.16 respectively.

In their review of studies on the question, Aparicio, Tejedor and Sanmartin (1982) indicated that, in general, a positive correlation has been found between grades and evaluations. However, the correlations obtained oscillated between -0.75 and 0.75. This lack of agreement is attributed to methodological problems.

According to Marsh (1984), the positive relationship between the average expected grade and ratings poses three hypotheses:

- The grading leniency hypothesis: instructors give higher grades and receive higher ratings than are deserved.
- The validity hypothesis: higher expected grades reflect more students learning and a positive correlation justifies the validity of student ratings.
- The student characteristics hypothesis: there are pre-existing predictive variables for students - such as previous subject interest - that affect student learning, grades and teaching effectiveness; thus, the effect of expected grades is spurious.

To conclude this point, we present the results of a study carried out at the Universidade da Coruña by De Salvador (1996). The sample in this study used 345 subjects corresponding to 16 degrees -11 faculties - and more than 20 questionnaires answered in each subject. The average rating of each of the three dimensions that make up the student evaluation of teaching (SET) questionnaire at this university was obtained in each subject, as were the grades given to student. The following is a summary of the results:

- A clear relationship was observed between grades given by instructors and stu-

dent ratings on the SET questionnaire. This leads the author to suggest a halo effect: high/low grades in a course would have a certain correlation with high/low student ratings for that subject. The correlations found between grades and the instrument measures were: the good instructor item, 0.391; teaching methodology factor, 0.350; interaction/rapport factor, 0.482, and evaluation factor, 0.496, all with significances of less than 0.001.

- This relationship is influenced more by “good grades” than by failures.

- The halo effect is consolidated by the congruence of the rapport, taking into account the questionnaire components: the evaluation system and rapport with students are the most influenced dimensions.

- The greatest influence is seen in the most advanced courses in subjects taught by a single instructor, especially in three-year diploma programmes.

- General courses show a quadratic trend of the effect with a highpoint in the third cycle: it increases moderately during the first cycle and decreases, also moderately, in the second cycle.

The author concludes that grades may distort SETs, in the understanding that the principle of theoretical coherence - good instructors result in good grades for their students - cannot be interpreted given the results obtained; “however, we consider that this study is limited to a single year of the questionnaire’s application and that in any case, it must be extended to successive years with a cross-sectional, but also longitudinal treatment” (De Salvador, 1996, 125).

Instructor variables

The following are the variables associated with instructors that were studied in research on SET questionnaires as possible modulators of responses.

Age, gender and physical appearance

A study by Goebel and Cashen (1979) found that students tend to give lower ratings to older and less attractive instructors. The gender effect was not statistically significant. The analysis of interactions showed that unattractive middle-aged women and unattractive older men tended to receive the lowest scores.

In this sense, Villa asserted that “the older and less attractive instructors start out at a disadvantage. This reflects societal stereotypes, since instructors are being evaluated on the basis of characteristics other than professional ones: in making judgements about their professional competence, the validity of student evaluations is seriously questioned, according to the authors.” (Villa, 1985b, 45)

Pozo, Reboloso and Fernández (2000) conducted a study that aimed to determine the characteristics that define an “ideal instructor”. To do so, they applied a semantically differential scale composed of 29 bipolar adjectives to 2,221 students at the University of Almeria. The results showed how an instructor’s “attractive or interesting” physical appearance is highly significant. This finding led the authors to hypothesise that emotional aspects have a substantial influence on student attitudes towards their instructors and consequently, on their conceptualisation of them.

Recently, the author of the present study carried out a qualitative study with qualitative interviews and discussion groups whose objective was similar to the one above. The analysis of content resulted in an emerging subcategory called “physical appearance” in relation to several students’ statements on the positive effect an instructor’s image can have on the global idea of the instructor. This subcategory was the one with the least associated verbal production, and therefore its presence in the general node structure was questioned because of its very low representativity (Casero, 2010).

Personality

The review and synthesis Feldman (1986) conducted of studies that relate student evaluations and personality traits or categories obtained from self-reports and hetero-evaluations - students, colleagues - show the following:

- Self-reporting: positive correlations only in positive self-image/self-esteem ($r = 0.30$) and energy/enthusiasm ($r = 0.27$). The average correlation between student ratings of effective teaching and each of the 12 remaining categories was 0.15 or less.

- Correlations were much higher when personality was rated by outside observers - colleagues and students - and the average of these among student ratings and the 14 personality traits ranged 0.30 to 0.60.

The author offers several possible explanations for these differences:

- The instructor's personality perceived by students may be affected by the same biases that affect evaluations of effective teaching.

- The personality inferred by colleagues may be based in part on information from students.

- The personality inferred by students and colleagues may be based in part on perceptions of effective teaching, rather than personality or vice-versa.

- The personality inferred from self-reports may be more or less valid or more or less biased than the personality inferred by students and colleagues. The personality inferred by students and perhaps by colleagues may be limited to a specific situational aspect, while personality measures based on self-reports are more general and do not focus on a specific context.

To Marsh (1987), Feldman's review suggests the existence of a relationship between student ratings of an instructor and at least several aspects of the instructor's personal-

ity, but does not indicate whether it is a valid source of influence or bias.

Diaz-Aguado (1987) believes the best way to consider the influence of the instructor's individual traits - as a personality type or cognitive style - on teaching evaluations lies in interaction with the student's same traits.

Teaching the course on several different occasions

This section seeks to determine whether the fact that an instructor has taught a course more than once drives up ratings. Marsh reached the following conclusions: "Correlations between student ratings of the same course taught by the same instructor on two different occasions were high, $r=0.71$. For each pair of courses, 341 pairs, the more favourably evaluated tended to be

- (1) the one in which students expected higher grades (and presumably learned more);

- (2) the one which students perceived to require the most work; and

- (3) the one which was taught after the instructor had already taught the course at least once before (and presumably improved as a consequence of this experience)."

In summary, the author states that "this study relates differences in the ratings of the same course taught by the same instructor on two different occasions with differences in the background characteristics of the two courses. For each pair of courses the most favourably evaluated correlated with difficulty/workload; higher expected grades - and presumably more learning; and with the instructor having already taught the course at least once before. These findings are not consistent with the hypothesis that these background characteristics bias student ratings, and they argue for alternative explanations." (Marsh, 1982 b, 496).

Volunteering to be evaluated

A study conducted by Howard and Bray (1979) found that a group of volunteers - instructors who continued to use student evaluations after a first course when they were required to do so - were rated higher than those who did not volunteer.

On the basis of this study, Cashin and Perin (1983) used the Instructional Development and Effectiveness Assessment (IDEA) instrument to analyse student evaluations from 13,063 classes in a range of academic fields and institutions. The classes were divided into three groups: volunteers, in which the decision to be evaluated was entirely the instructor's; intermediates, in which evaluation was required, but the instructor chose the class; and non-volunteers, in which instructor evaluation was required by class. The differences between the three groups were statistically significant for 26 of the 39 items on the IDEA - probably because of the large number of cases. Based on an omega-squared analysis, none of these differences have practical significance - the sizes of the effect did not even reach 1% of the variance. It was concluded that the voluntary nature of evaluation need not be taken into account when using large, multi-institutional, comparative data collection instruments.

It may be appropriate to question the hypothesis that the higher the effectiveness ratings, the more concern for improving and logically, greater interest in volunteering to ascertain the students' opinions of teaching effectiveness.

Rank and experience

Another issue studied has been the teacher's rank as a possible influence on SETs. A number of authors have presented studies on this topic, one of the most outstanding of which was conducted by Feldman, who indicated that teaching assistants generally receive lower ratings than other instructors in most of the SEEQ dimensions and on some items, although they may receive higher rat-

ings in the "individual rapport" and "group interaction" dimensions. Most of the studies this author reviewed found that teacher rank had no significant effect on overall evaluations and that the significant relationships tend to be positive. Rank is also not significantly related to the dimensions of the ratings in most studies; when they are positive, it is most likely to be in "instructor knowledge" and "breadth of coverage", while negative relations appear in "encouragement of discussion" and "openness and concern for students".

Regarding experience, and closely related with age, Braskamp et al. (1985) found that student evaluations may rise during the first 10 years of teaching, but fall little by little after that.

The Dr. Fox paradigm

The Dr. Fox effect, also known as "educational seduction", is defined as the predominant influence of the instructor's expressiveness. This phenomenon is usually associated with the notion that a teacher's enthusiasm may entice students into rating his or her performance favourably, even when the instruction lacks relevant content.

The original study on the Dr. Fox effect was conducted by Naftulin, Ware and Donnelly in 1973. The authors hired a professional actor to give a series of conferences to three groups of medical students and instructors; the actor was presented as "Dr. Myron L. Fox". The actor had been trained beforehand to use neologisms, ambiguities, contradictions and meaningless phrases in his lectures. In short, Dr. Fox gave an entertaining lecture series with little or no content. Both the students and the instructors assessed the instructor's performance favourably.

Ware and Williams (1975, 1977) and Williams and Ware (1976, 1977) developed the standard Dr. Fox paradigm with a series of six explanations, all presented by a professional actor on videotape. Each explanation represented one of three levels of course con-

tent and one of two levels of expository expressiveness. The students completed an evaluation questionnaire with several items and a performance test. The conclusions were that differences in expressiveness consistently explain much more variance in student ratings than differences in content.

In a re-analysis of previous studies, Marsh and Ware (1982) concluded that manipulating instructor expressiveness only affects the instructor's "enthusiasm" ratings, the factor most logically connected with expressiveness, and that breadth of content significantly affected ratings of the "instructor knowledge" and "organization/clarity" factors, the ones most logically related to expressiveness. Marsh states that "an effect that has been interpreted as a bias to SETs seems more appropriately interpreted as support for their validity with respect to one component of effective teaching" (Marsh, 1987, 333).

Abrami, Leventhal and Perry (1982) conducted a meta-analysis of all known studies on the subject and concluded that above all, manipulations of expressiveness have a substantial effect on student ratings and a small effect on performance, while manipulations of content have a substantial effect on performance and a small effect on ratings. They concluded that, although expressiveness interacted with the manipulation of content and a group of other variables examined in the Dr. Fox studies, none of the interactions accounted for more than 5% of the variance in SETs.

Background variables

This group of variables includes those that cannot be directly associated with the student or the instructor, such as those related to the subject's characteristics: time of day the class is taught, type and volume of workload/difficulty, class size and time of year when students respond to the teaching evaluation questionnaire.

Time of day when the class is taught

Some instructors believe that classes given late in the morning or late in the afternoon may be rated lower, because many students have already left. It is true that students leave, but it is not true that the evaluations are biased by this factor and so do not fit reality. At the most, in these cases there would be fewer students with a formed judgment allowing them to evaluate the instructor. The truth is that there is a dearth of studies in this regard, as indicated by Aparicio, Tejedor and Sanmartín (1982).

Cranton and Smith (1986) studied the differences between day and evening courses in a multivariate analysis and found very little difference. However, in a separate study - the univariate analysis - no significant differences in teacher ratings were found, although differences were found in ratings of the amount of learning, the importance thereof and the overall evaluation.

Type of subject

At this point, we reflect on the considerations and studies of the subject's importance in the curriculum, its required or elective nature and its nature as a discipline - science, literature, etc.

As indicated by Aparicio, Tejedor and Sanmartín (1982), no relationship was found between the subject's importance in study programmes and student ratings. These authors note that the results of studies that relate evaluations and a subject's elective/obligatory nature are contradictory. Several studies claim that students tend to give instructors of electives higher ratings, while others found no differences.

On another note, several authors have linked ratings with type of academic discipline. In a review of studies comparing evaluations across disciplines, Feldman (1978) found that that ratings were somewhat higher than average in English, the humanities, arts, language and education and somewhat lower in social sciences, physical sciences, mathe-

mathematics and engineering, with biological sciences being around the average.

In a large-scale study of 100 institutions, Centra and Creech (1976) found that ratings were higher in the humanities and lower in sciences.

Based on the Biglan's classification of academic fields (1973) - a) hard/soft, b) applied/pure, and c) life/nonlife - Neumann and Neumann (1985) indicated that ratings might be higher in soft, pure and nonlife disciplines; thus, comparisons should be made between instructors in like areas. However, this study was conducted at only one institution, so its generality should be tested.

The results of these studies may have a greater significance and importance for summative rather than formative evaluations. The disciplines Feldman indicates as the highest rated seem to lend themselves to more instructor-student interaction, the use of more active methods and ultimately, student participation.

Student questionnaires were administered in a study conducted by Garcia Valcárcel (1989) at the University of Cantabria and two distinctly different models appeared after a cluster analysis of the average responses per item/subject:

- Informative model: 84 subjects. It is characterised by the following elements: memorisation of knowledge, lecture mode presentation, use of a single textbook and use of traditional exams. The instructor informs, the student assimilates, the student demonstrates what has been assimilated on a test and the instructor grades.

- Communicative or formative model: 62 subjects. This model obtained high ratings on items referring to objectives, motivation, interaction and occasionally, use of teaching techniques other than lectures.

Subject workload/difficulty

Subject workload/difficulty is another aspect that has been considered. Among the reasons that attention should be paid to the presence of an effect related to subject workload/difficulty are those reported by Ryan, Anderson, and Birchler (1980) when they assert that the introduction of SETs in an institution leads to easier courses with lighter workloads, in the belief that this might lead to higher SETs.

Research on this aspect in questionnaires coincides in that difficult courses with high workloads are associated positively with more favourable evaluations, other aspects being equal. To investigate this question, the same course taught by the same instructor at different times is usually studied. However, although correlation is lower in instructor self-evaluations, it still heads in the same direction. Thus, it cannot be said that the workload/difficulty of a course biases student evaluations (Marsh, 1987).

Class Size

In general, instructors believe that colleagues who teach small classes are rated higher than teachers with large classes. The studies conducted on the subject have not found any relationship in most cases, while a quadratic correlation has appeared in a few. However, the results are somewhat contradictory.

Aparicio, Tejedor and Sanmartin (1982) cited studies in which small class ratings are higher than larger groups'.

Aleamoni and Graham (1974), among others, did not find a significant relationship between class size and student ratings of instructional quality.

Marsh and his colleagues found that class size moderately correlates with "group interaction" and "individual rapport"- negative correlations of up to 0.30 - but not with other dimensions or global evaluations of the course instructor. The author states that "the

specificity of class size effect to dimensions most logically related to this variable, and the similarity of findings based on SETs and faculty self-evaluations argue that this effect is not a “bias” to SETs; rather, class size does have moderate effects on the aspects of effective teaching, primarily group interaction and individual rapport to which it is most logically related, and these effects are accurately reflected in the SETs.” (Marsh, 1987, 314)

Feldman (1978) conducted an extensive review and found results consistent with Marsh’s on the SEEQ.

Mateo and Fernández (1996) conducted a study of 5,959 different-sized classes that ranged from 3 to 498 students per class at the University Complutense of Madrid. They classified the sizes according to the following five categories:

- Very small: between 3 and 9 students
- Small: 10 to 29 students
- Medium: 30 to 59 students
- Large: 60 to 149 students
- Very large: more than 149 students

The instrument used was the University’s own – the abridged version of the CUTEQ-R – which analysed the two dimensions of “competence” and “motivation and interaction skill”.

The statistical treatment consisted in analysing the unifactorial variance of each dimension under the factor described above. The results showed statistical significance in both dimensions, reaching eta squared indices of 0.0299 in “competence” and 0.0580 in “motivation and interaction skills”. The following table shows the averages obtained.

Table 1. Averages obtained in the two dimensions analysed according to class size

Size/Dimension	Competence	Motivation and interactional skills
Very small	5.14	5.82
Small	4.80	4.91
Medium	4.67	4.74
Large	4.87	4.75
Very large	4.58	4.50

Reproduced from Mateo and Fernández (1996, p. 776)

The authors concluded that the results support the presence of some kind of effect of class size on students ratings, citing similarities with Feldman’s and Marsh’s results, and noted a differential influence on the dimensions of the teaching evaluation. This effect did not surpass 5% of the variance explained in most cases.

Salvador (1990) proposes a distinction in the nature of the relationship between class size and student ratings. He indicates that we might speak of “bias” when higher ratings are obtained simply for teaching a small class. It is not the same when size is an integral element in effective teaching, i.e., that more interaction and contact with students results in more effective teaching and consequently, higher ratings.

Point in the course when the questionnaire is administered

In this last point, we address the possible influence of the point in the course when evaluations are completed.

Feldman (1979) reported that the mid-year and end-of-course evaluations tend to be similar.

Braskamp et al. (1985) suggested that evaluations administered during the final exam are lower and those administered mid-course less reliable if students can be identified.

Marsh and Overall (1980) collected ratings in the middle of the course and during the last week of the course in their study of multi-section validity. Both were highly corre-

lated, but the validity coefficients of mid-course evaluations were substantially lower.

Braskamp et al. (1985) recommended that evaluations be administered during the last two weeks of a course. This is more or less the customary practice at Spanish universities today.

Not to conclude

The review of research on the “biases” that may affect student evaluations of instructional quality has revealed the existence of several variables that present correlations in some cases that - although statistically significant, the logical result of working with large-sized samples - do not constitute a substantial effect that undermines the validity of the instruments analysed, as several authors have noted (Marsh, 2007).

The contradictory results detected in several studies can probably be attributed to a mismatch in the methods, scales, units of measurement used and in some cases, questionable possibilities of representativeness and generality. The lack of methodological rigor is joined by errors in interpreting some of the results. It seems logical that a senior professor teaches more effectively and thus receives better student ratings - Is teacher rank thus a bias? - or that large class sizes lead to lower ratings - Is class size as a source of error or teaching conditions as an explanatory factor of the quality thereof? These examples, among others, illustrate what Feldman (1997) called the myths and half truths in teacher evaluation.

After fifty years of research, the literature shows, though still with little impact, that measurements of teaching evaluations by students are reliable and valid. Few objects of study in the field of evaluation have been examined so often, and as a result of this obsessive scrutiny, they enjoy more guarantees than many other widely-used instruments. However, research in recent decades has not been especially helpful in resolving the loose ends detected in the 1980s, the

golden decade of SET research. The most recent scientific output is oriented towards psychometric refinement in the Anglo-Saxon case, while the Spanish literature of the past few decades seems more concerned about the purpose of summative and/or formative evaluations without coming to a consensus on the matter: while authors such as De Miguel (1991) argue for the complementarity of the two purposes, others such as Escudero (1993) are opposed to this view. In turn, Meliá (1993) and Mateo (2000) propose mixing processes based on context and consensus and Apodaca and Grad (2002) stress that the objective of evaluation is what determines the uni- or multi-dimensional view of the construct. In this sense, as has been pointed out elsewhere (Casero, 2008, 2010), the multidimensional model receives the majority support.

Recapitulating, it is interesting how such a scrutinised, well-defined object of study continues to generate controversy over decades. The question is why? Perhaps part of the answer, if not all, can be explained by the point evaluation impacts: competent teaching. This seems to be an extremely delicate matter, especially if judgement of that competence lies in student hands. The teaching evaluation system is made up of a series of stakeholders, including students, however this has been the only source of information used at many Spanish universities. This, together with personal prejudices, fear, distrust, ignorance and administrative convenience, among other reasons, has sustained the unproductive debate on the reliability and validity of teacher evaluations, thus ignoring fifty years of research findings (Theall and Feldman, 2007). The problem seems to originate in the lack of courage, mainly in institutions. Universities must tackle the creation of global evaluation systems, develop teacher training programmes consistent with high standards of teaching and foster a culture of quality. Perhaps thus, with institutional involvement, better research can be carried out that, in addition to clearly improving the

Spanish university system, also leads to the absolute banishment of myths and half truths.

References

- Abrami, P.C., Leventhal, L. and Perry, R.P. (1982). Educational seduction. *Review of Educational Research*, 52, 444-464.
- Abrami, P.C. and D'Apollonia, S. (1990). The dimensionality of ratings and their use in personnel decisions. In M. Theall and J. Franklin (Comps.), *Student ratings of instruction: Issues for improving practice. New directions for teaching and learning*, 43 (pp. 97-111). San Francisco: Jossey-Bass.
- Aleamoni, L.M., Yammer, M. and Mahan, J. M. (1972). Teacher folklore and sensitivity of a course evaluation questionnaire. *Psychological Reports*, 31, 607-614.
- Aleamoni, L.M. and Hexner, PHZ (1980). A review of the research on student evaluation and a report on the effect of different sets of instructions on student course and instructor evaluation. *Instructional Science*, 9 (1), 67-84.
- Aparicio, J.J., Tejedor, F.J. and Sanmartín, R. (1982). *La enseñanza universitaria vista por sus alumnos: Un estudio para la evaluación de los cursos de la enseñanza superior*. Madrid: ICE/ Universidad Autónoma de Madrid.
- Apodaca, P. and Rodríguez, M. (1999). La opinión de los alumnos en la evaluación de la calidad docente: posibilidades, limitaciones y estructura dimensional de sus indicadores. In *Indicadores en la Universidad: información y decisiones* (pp. 311-327). Madrid: Consejo de Universidades/ Ministerio de Educación y Cultura.
- Barke, C.R., Tollefson, N. and Tracy, D.B. (1983). Relationship between course entry attitudes and end-of-course ratings. *Journal of Educational Psychology*, 75 (1), 75-85.
- Biglan, A. (1973). The Characteristics of Subject Matter in Different Academic Areas. *Journal of Applied Psychology*, 57 (3), 195-203.
- Braskamp, L.A. and Ory, J.C. (1985). *Evaluating teaching effectiveness: A practical guide*. Beverly Hills: Sage.
- Casero, A. (2008). Propuesta de un cuestionario de evaluación de la calidad docente universitaria consensuado entre alumnos y profesores. *Revista de Investigación Educativa*, 26, 25-44.
- Casero, A. (2010). ¿Cómo es el buen profesor universitario según el alumnado? *Revista Española de Pedagogía*, 246, 223-242.
- Cashin, W.E. and Perrin, B.M. (1983). Do college teachers who voluntarily have courses evaluated receive higher student ratings? *Journal of Educational Psychology*, 75 (4), 595-602.
- Castro, J.J. (1996). *Factores moduladores de la evaluación del profesor*. Doctoral thesis. Universidad de la Laguna, Tenerife.
- Centra, J.A. and Creech, F.R. (1976). *The Relationship Between Students, Teachers, and Course Characteristics and Student Ratings of Teacher Effectiveness*. Princeton, N.J.: Educational Testing Service.
- Cranton, P.A. and Smith, R.A. (1986). A new look at the effect of course characteristics on student ratings of instruction. *American Educational Research Journal*, 23 (1), 117-128.
- De Salvador, X. (1996). Sobre la evaluación de la actividad docente del profesorado universitario: ¿Está mediatizada la valoración de los alumnos por las calificaciones? *Revista Española de Pedagogía*, 203, 107-128.
- Díaz-Aguado, M.J. (1987). La percepción del profesor por el alumno: expectativas y actitudes. In J. Beltrán (Ed.), *Psicología Educativa: Vol. 1*. Madrid: UNED.
- Etopa, M.P. (2003). *Evaluación del profesorado de la Universidad de Las Palmas de Gran Canaria: Repercusiones del paso de una evaluación formativa a una evaluación sumativa*. Doctoral thesis. Universidad de Las Palmas de Gran Canaria, Canary Islands.
- Feldman, K.A. (1978). Course characteristics and college students' ratings of their teachers and courses: What we know and what we don't. *Research in Higher Education*, 9, 199-242.

- Feldman, K.A. (1979). The significance of circumstances for college students' ratings of their teachers and courses: a review and analysis. *Research in Higher Education*, 10, 149-172.
- Feldman, K.A. (1983). Seniority and experience of college teachers as related to evaluations they receive from students. *Research in Higher Education*, 18, 3-124.
- Feldman, K.A. (1986). The perceived instructional effectiveness of college teachers as related to their personality and attitudinal characteristics: a review and synthesis. *Research in Higher Education*, 24, 129 - 213.
- Feldman, K.A. (1997). Identifying exemplary teachers and teaching: Evidence from student ratings. In R.P. Perry and J.C. Smart (eds.), *Effective Teaching in Higher Education: Research and Practice* (pp.368- 395). New York: Agathon Press.
- Feldman, R.S. and Prohaska, T. (1979). The student as Pygmalion: Effect of student's expectancy on the teacher. *Journal of Educational Psychology*, 71, 485-493.
- Fernández, J. and Mateo, M.A. (1997). Student and faculty gender in ratings of university teaching quality. *Sex Roles*, 37 (11/12), 997- 1003.
- García Valcarcel, A. (1989, June). *Modelos didácticos en la Universidad de Cantabria* (Inf. de investigación). Universidad de Cantabria.
- Goebel, B.L. and Cashen, V.M. (1979). Age, sex and attractiveness as a factor in student rating of teachers. *Journal of Educational Psychology*, 71, 646-653.
- Gough, H.G. (1979). A creative personality scale for The Adjective Check List. *Journal of Personality and Social Psychology*, 37, 1398-1405.
- Hernández, P. (1991). *Psicología de la Educación: Corrientes Actuales y Teorías Aplicadas*. México: Trillas.
- Howard, G.S. and Bray, J.H. (1979). Use of norm groups to adjust student ratings of instruction: A warning. *Journal of Educational Psychology*, 71, 58-63.
- Marsh, H.W. (1980). The influence of student, course, and instructor characteristics in evaluation of university teaching. *American Educational Research Journal*, 17 (1), 219-237.
- Marsh, H.W. (1982). Factors affecting student's evaluations of the same course taught by the same instructor on different occasions. *American Educational Research Journal*, 19 (4), 485-497.
- Marsh, H.W. (1984). Student's evaluations of university teaching: dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology*, 76, 707-754.
- Marsh, H.W. (1987). Students' Evaluation of University Teaching: Research Findings, Methodological Issues, and Directions for Future Research, *International Journal of Educational Research*, 11 (3), 255-388.
- Marsh, H.W. (2007). Students' Evaluations of University Teaching: Dimensionality, Reliability, Validity, Potential Biases and Usefulness. In R.P. Perry and J.C. Smart (eds.), *The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective*, (pp.319-383). Netherlands: Springer.
- Marsh, H.W. and Overall, J.U. (1980). Validity of students' evaluations of teaching effectiveness: Cognitive and affective criteria. *Journal of Educational Psychology*, 72, 468-475.
- Marsh, H.W. and Ware, J. (1982). Effects of expressiveness, content coverage and incentive on multidimensional student rating scales. *Journal of Educational Psychology*, 74, 107-116.
- Mateo, M.A. and Fernández, J. (1993). Dimensiones de la calidad de la enseñanza universitaria. *Psicothema*, 5 (2), 265-275.
- Naftulin, D.H., Ware, J.E. and Donnelly, F.A. (1973). The Doctor Fox lecture: A paradigm of educational seduction. *Journal of Medical Education*, 48, 630-635.
- Neumann, L. and Neumann, Y. (1985). Determinants of students' instructional evaluation: A comparison of four levels of academic areas. *Journal of Educational Research*, 78 (3), 152-158.

Pozo, C., Reboloso, E. and Fernández, B. (2000). The 'Ideal Teacher'. Implications for student evaluation of teacher effectiveness. *Assessment & Evaluation in Higher Education*, 25 (3), 253-263.

Ryan, J.J., Anderson, J.A. and Birchler, A.B. (1980). Student evaluation: The faculty respond. *Research in Higher Education*, 12, 317-333.

Salvador, L. (1990). *Los docentes universitarios exitosos desde la perspectiva del alumno: su caracterización psicopedagógica*. Doctoral thesis. Universidad de Salamanca.

Theall, M. and Feldman, K.A. (2007). Commentary and update on Feldman's (1997) "Identifying exemplary teachers and teaching: Evidence from student ratings". In R.P. Perry and J.C. Smart (eds.), *The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective*, (pp.130-143). Netherlands: Springer.

Villa, A. (1985). La evaluación del profesor: perspectivas y resultados. *Revista de Educación*, 277, 55-93.

Ware, J.E. and Williams, R.G. (1975). The Doctor Fox Effect: A study of lecturer effectiveness and ratings of instruction. *Journal of Medical Education*, 50, 149-156.


Ware, J.E. and Williams, R.G. (1977). Discriminate analysis of student ratings as a means for identifying lecturers who differ in enthusiasm or information giving. *Educational and Psychological Measurement*, 37 (3), 627-639.

Williams, R.G. and Ware, J.E. (1976). Validity of student ratings of instruction under different incentive conditions: a further study of the Dr. Fox effect. *Journal of Educational Psychology*, 68 (1), 48-56.

Williams, R.G. and Ware, J.E. (1977). An extended visit with Doctor Fox: validity of student ratings after repeated exposures to a lecturer. *American Educational Research Journal*, 14 (4), 449-457.

ABOUT THE AUTHORS / SOBRE LOS AUTORES

Casero, Antonio (a.casero@uib.es). Bachelor's in Psychology and Doctor of Education Sciences. Since 1998, he has been a lecturer in the area of Didactics and Scholastic Organisation at the *University of the Balearic Islands*' Faculty of Education (Spain). He is a member of the Education and Citizenship research group (EIC) and his research activities revolve around the fields of quality university teaching, teacher training, cyberplagiarism and the socio-laboural insertion of groups at risk of social exclusion. His postal address is: Departamento de Pedagogía Aplicada y Psicología de la Educación. Universitat de les Illes Balears. Crta de Valldemossa, km 7.5. 07122 Palma de Mallorca (Islas Baleares, España). [Buscar otros artículos de este autor en Google Académico /](#)

[Find other articles by this author in Scholar Google](#) 

ARTICLE RECORD / FICHA DEL ARTÍCULO

Reference / Referencia	Casero, Antonio (2010). <i>Modulating factors in the perception of teaching quality</i> . <i>RELIEVE</i> , v. 16, n. 2. http://www.uv.es/RELIEVE/v16n2/RELIEVEv16n2_3eng.htm .
Title / Título	<i>Modulating factors in the perception of teaching quality</i> . [Factores moduladores de la percepción de la calidad docente].
Authors / Autores	Casero, Antonio
Review / Revista	RELIEVE (Revista ELeCtrónica de Investigación y EValuación Educativa / <i>E-Journal of Educational Research, Assessment and Evaluation</i>), v. 16, n. 1.
ISSN	1134-4032
Publication date / Fecha de publicación	2010 (Reception Date : 2009 September 25; Approval Date : 2010 September 30; Publication Date : 2010 September 30).
Abstract / Resumen	<p><i>This article examines the existing research findings in relation to the research of the factors that are presumably involved in university students' evaluations of the teaching quality they receive, possibly threatening the validity of the construct. The review is organized around the three implicated sources in the perceptive process of the student, being the student himself, the professor, and the context. The review concludes that, despite some minor effects, the factors under analysis do not represent a substantial threat to the system's validity. The unproductive debate that has gone on for decades about guarantees in the evaluation of teaching staff has largely been used to cover up the absence of a desire to institutionalize this kind of evaluation system.</i></p> <p>El presente artículo tiene como objetivo presentar una revisión sobre los hallazgos existentes en relación a la investigación de los factores supuestamente implicados en la valoración que el alumnado universitario realiza sobre la tarea docente de sus profesores, pudiendo suponer una amenaza a la validez de dicho constructo. La revisión se presenta organizada en torno a las tres fuentes implicadas en el proceso perceptivo del alumno, siendo éstas el propio alumno, el profesor y el contexto. La revisión concluye que, a pesar de efectos menores, los factores analizados no representan una amenaza substancial a la validez del sistema. El debate infructuoso que se ha producido durante décadas sobre las garantías en la valoración del personal docente ha sido usado en gran parte para encubrir la falta de deseo de institucionalizar este tipo de sistemas de evaluación.</p>
Keywords / Descriptores	<i>Student evaluation of Teacher Performance, Teaching Quality, Effective Teaching.</i> Evaluación del desempeño docente, Calidad docente, Efectividad docente
Institution / Institución	Departamento de Pedagogía Aplicada y Psicología de la Educación. Universitat de les Illes Balears (España).
Publication site / Dirección	http://www.uv.es/RELIEVE
Language / Idioma	Español and English version (Title, abstract and keywords in English and Español)

RELIEVE

Revista ELeCtrónica de Investigación y EValuación Educativa
E-Journal of Educational Research, Assessment and Evaluation

[ISSN: 1134-4032]

© Copyright, RELIEVE. Reproduction and distribution of this articles it is authorized if the content is no modified and their origin is indicated (RELIEVE Journal, volume, number and electronic address of the document).

© Copyright, RELIEVE. Se autoriza la reproducción y distribución de este artículo siempre que no se modifique el contenido y se indique su origen (RELIEVE, volumen, número y dirección electrónica del documento).