

The dimensionality of student ratings of teaching: integration of uni- and multidimensional models

Pedro Apodaca^{a*} and Hector Grad^b

^aUniversidad País Vasco, Spain; ^bUniversidad Autónoma de Madrid, Spain

The debate on the unidimensionality *versus* multidimensionality of students' ratings of teaching performance, and the relationship of this dilemma to summative and formative uses of these instruments is reviewed, with special attention to the implications of their development and application in the Spanish university system. A comprehensive theoretical framework, stressing the relation between effective teaching theory and the expected construct of teaching performance in students' view, is advanced. The structure of teaching competencies is analysed by means of confirmatory factor analysis (CFA) and similarity structure analysis (SSA) in two ratings forms, of 13 and 34 items. The forms were applied to large student samples from two Spanish universities, rating the performance of individual teachers. CFA and SSA show a common latent structure in the two samples, and support the conceptual distinction between aspects of teaching competence. The results suggest that the structure of the students' ratings can be interpreted as multidimensional as much as unidimensional. The reason for this apparent paradox is that some specific dimensions are more central (*versus* peripheral) in the construct of teaching skills. This structure is consistent with the conceptual framework, and would explain the ambiguity between unidimensional and multidimensional solutions of CFA outcomes. The implications of these results for the orientation of future research on the dimensionality of students' ratings and for their application in formative and summative evaluation of teaching are discussed.

Introduction

Students' ratings of teaching performance have a comparatively short history in Spain, since they were introduced after the democratization of Spanish universities in the early 1980s. Spanish universities, public in the main, shifted then from an elitist to a democratic model (Quintanilla, 1998). This modernization implied strong teaching and research demands for academic institutions and staff, since it combined

*Corresponding author. E. U Magisterio, Universidad País Vasco, c/Ramón y Cajal 72, 48014-Bilbao, Spain. Email: pedro.apodaca@ehu.es

the opening of higher education to students of broader social and educational backgrounds and a renewed stress on research in the context of strong structural constraints. In the teaching domain, this change led to a mass educational system, with very large and diverse class groups in most of the undergraduate studies. This situation reinforced traditional teaching customs, leaving conventional lecturing oriented to 'knowledge transmission' (Kember & Gow, 1994) as the main teaching method.

In this context, student surveys of teaching performance were launched, soon becoming an evaluation standard for all Spanish public universities. Their implementation resulted from both the cultural adaptation of American instruments and the development of indigenous instruments – mainly for formative purposes (see Aparicio *et al.*, 1982, for a pioneering study). Since these surveys aimed mainly at improving individual skills, there is still a need to enhance their formative function from the perspective of the professional development of academic staff (de Miguel, 2003; Rodríguez, 2003).

With the consolidation of this evaluation method, however, the same instruments and procedures began to be used for summative purposes. Recent laws regulating the Spanish university system (Constitutional Law 6/2001, of Universities) have gone further, in establishing evaluation requirements for academic staff and linking salary increments to quality and productivity assessments. Thus, there is a need to review the appropriateness of these instruments and procedures for such new functions, as well as to clarify some theoretical issues with significant implications for their summative use.

Given the probable increasing role of student surveys of teaching performance for staff evaluation, the proper use of these surveys will require paying special attention to the construct validity of the rating forms. The unidimensional or multidimensional structure of the construct of teaching competence from the student perspective merits a central place in this regard.

While the multidimensionality of this construct is hardly challenged from the point of view of effective teaching theory and practice, the dimensionality of rating forms of this competence, firstly, depends on students' particular perceptions. Furthermore, the dimensionality debate is affected by the debate about the formative and the summative goals of teaching evaluation. These goals impose contradictory requirements, from the practical point of view, for the evaluation design: whilst formative goals require precise and specific feedback as a guide to the improvement of teaching performance, summative goals need an overall parsimony that facilitates the survey application and subsequent decision-making based on the obtained information. The solution to this dilemma has direct implications for the appropriate design of instruments and information feedback to the stakeholders of the evaluation process. The dilemma is often reduced to the choice between multidimensional instruments with a high level of specificity and the reporting of scores for the different specific dimensions, or the gathering of global ratings and/or the report of a single general score of 'teaching competence' (or, more generally, of overall satisfaction with the teaching).

There is abundant research on the dimensionality of student ratings of teaching performance in the English-speaking context. The studies of Marsh (Marsh, 1991; Marsh & Hocevar, 1984, 1991), Abrami and d'Apollonia (Abrami, 1989; Abrami & d'Apollonia, 1990; Abrami *et al.*, 1996; d'Apollonia & Abrami, 1997), and Feldman (Feldman, 1976, 1996) are classical references in the field. In spite of this abundance of studies, which includes large reviews and meta-analyses, the multidimensional *versus* unidimensional debate remains open.

On the one hand, the multidimensional approach seems more appropriate for formative purposes. The precise assessment of the specific aspects of teaching performance allows the identification of profiles of both teaching quality and improvement needs. From this point of view, a single overall measure of performance would have no diagnostic value, and implies a great loss of useful information for many purposes (Marsh, 1991).

On the other hand, Abrami and d'Apollonia (1990) defended the unidimensional approach and, specifically, the use of a single overall measure, based either on overall rating items or on a weighted average of items reflecting specific teaching behaviours, for summative purposes. D'Apollonia and Abrami (1997) added interesting arguments in favour of unidimensional ratings. Firstly, on content validity grounds, some dimensions of teaching competence would be affected by factors beyond the teacher's domain. Secondly, on predictive validity grounds, there is abundant evidence that the correlation between students' ratings and actual learning is higher for overall items than for those reflecting specific dimensions (as the classical meta-analysis of Cohen, 1981, pointed out). Thirdly, on operational grounds, the use of multidimensional criteria is too complex for decision-makers lacking expertise in teaching-learning and evaluation processes. They may inappropriately weight and average dimensions and make decontextualized comparative judgements. In summary, d'Apollonia and Abrami (1997) recommended the exclusive use of overall ratings of teacher effectiveness for summative purposes.

Regarding structural validity, Abrami *et al.* (1996) rejected the generalizability of the multidimensional structure across different rating forms, student or group characteristics, teaching methodology and so on. Nonetheless, Marsh and Hocevar (1984, 1991) reported results in support of this generalizability across different teaching aspects, such as discipline, instructor level and course level.

Cross-cultural research would contribute, in this regard, to the discussion of the dimensionality question making feasible the contrast of models and outcomes obtained in the English-speaking context with different instruments and populations. Some research has actually addressed this question in the Spanish context. Among the published studies, the work of Tejedor and colleagues (Tejedor, 1993; Tejedor *et al.*, 1987, 1988) is an unavoidable reference, because of its impact in the early stages of the development of students' rating forms and their implementation in Spanish universities. Among the recent studies, García Ramos (1997) introduced the use of confirmatory factor analysis (CFA), thus bringing the methodology of Spanish research up to date. Nevertheless, these studies have not solved the questions left open by English-context research.

One reason for this shared failure is that the debate on the dimensionality of teaching ratings has been strongly constrained by the use of factor analysis as the hegemonic statistical technique for structural analyses. The limitations of this technique for generating theory and identifying substantive aspects of a construct have been depicted in general (for instance by Guttman, 1982a, 1982b; Elejabarrieta & Perera, 1989), as well as in their application to solve this dimensionality question (as Feldman, 1976, warned nearly 30 years ago). Pointing out this problem, d'Apollonia & Abrami (1997) suggested that the specific factors found in multidimensional solutions may be artefacts of the statistical technique applied.

Though non-parametric structural analyses like Guttman's smallest space analysis (1968 – SSA, lately renamed as similarity structure analysis) would avoid most of factor analysis shortcomings, these techniques have scarcely been applied in this field of research. However, the lack of assumptions about variable distributions, their ability for representing linear (dimensional) and non-linear latent structures (such as circular order), and for formalizing systematic hypotheses (thanks to their application combined with facet theory – Canter, 1985; Dancer, 1990; Borg & Shye, 1995; Guttman & Greenbaum, 1998) suggest that these techniques can clarify some conceptual questions, bring an operative integration of uni- and multidimensional approaches and, then, support the use of student ratings for both summative and formative evaluations of teaching.

Thus, following previous research by Apodaca (Apodaca & Rodríguez, 1999; Apodaca & Grad, 2002), striving to integrate uni- and multidimensional approaches, the present study will contribute to this debate, both by a conceptual reassessment of the dimensionality of teaching performance and by an innovative triangulation of linear CFA and SSA structural outcomes. Furthermore, the reliability of the results will be tested by replication of the structural analyses on the rating forms used in two Spanish universities. The multiplicity of both forms and analysis techniques differentiates this study from Abrami *et al.* (1996), which covered studies applying diverse rating forms, but with factor analysis as the single predominant technique.

We will now review briefly the literature on the dimensionality of student rating forms of teaching performance, both from the conceptual and the applied perspectives. The report of the structural analyses will be also preceded by a discussion of the features and limitations of the parametric and non-parametric statistical techniques applied in the field.

Dimensions of teaching competence: a conceptual approach

The design of evaluation instruments for teaching performance at higher education should take into account the aspects or dimensions that the theory of the teaching-learning process identifies as necessary conditions for effective adult learning.

From a student learning approach, Entwistle (1987) suggested a heuristic model for analysing the teaching-learning process at higher education. The model focuses on the factors that may influence the learning strategies, processes and outcomes of the student (antecedents of learning strategies), the teaching style (teaching methods,

resources, and skills, relationship with the students) and the institutional context (including learning materials, evaluation procedures, overall study load, teaching feedback, professionalization of teaching and freedom to learn). This model highlights the diversity and complexity of the factors that should be taken into account for the assessment of the teaching-learning process. This assessment would require a multidimensional approach, as well as the specification of the appropriate sources and procedures for gathering information on the different aspects of the process.

Focusing on teaching, gathering valid and reliable information on each aspect and stage quality will require a specific combination of quantitative and qualitative methods (statistical indicators, observation, opinion surveys, in-depth interviews, content analysis, etc.) and sources (present and past students, expert peers in the taught field and in teaching methods, institutional managers, etc.) (see, for example, de Miguel, 1991). The relevance of each source will depend on both its relation to the teaching (direct observer or indirect source) and its competence for a valid and reliable evaluation of the considered aspect.

It is important to point out that, even if they are privileged observers, the students' opinions represent a partial and biased view of the 'teaching competence' construct (Abrami *et al.*, 1996, p. 223). Given their very position in the teaching-learning process, students cannot be a valid and reliable source of information on those aspects of teaching they cannot observe systematically, or in which particular motives or conflicts of interest may clearly bias their perceptions and evaluations. Thus, the use of students' ratings must necessarily form part of a wider system of teaching evaluation, which involves other criteria, information sources, evaluation instruments and activities (Seldin, 1995; de Miguel, 1998; Gillespie *et al.*, 2002).

Specifically, students may be appropriate information sources on the majority of lecturing aspects. Likewise, they can evaluate the observable aspects of planning, but they would be questionable judges of the scientific content of programmes. Finally, although the content and procedures of learning evaluation are the factor that most influences study strategies and outcomes (Entwistle & Ramsdem, 1983), the relationship of this evaluation to students' academic recognition and progress causes a conflict of interests that may bias their perceptions and judgements of this aspect of teaching. This conflict would restrict the validity of students' information on the quality of learning evaluation, unless it was possible to control the effect of their grade expectations on their opinions about this teaching aspect.

The evaluation of teaching quality based on students' ratings should, therefore, take into account the principles for effective teaching stemming from learning theory, as well as the aforementioned conditions for the design of evaluation instruments. However, there is no comprehensive and accepted model in this domain. From a logical point of view, any comprehensive evaluation should cover the three basic stages of teaching: planning, performance and evaluation. Thus, from a perspective of academic staff training, Newble and Cannon (1995) proposed 'organization', 'instruction', 'evaluation', 'relationships' and 'subject mastery' as the most important aspects of teaching.

These aspects were synthesized by Lowman (1984) in a bi-dimensional model of 'intellectual stimulation' (explaining with clarity, subject mastery and ability to

modulate the voice) and ‘rapport’ (interpersonal skills, as well as stimulation of student motivation and independent learning).

Drawing on the classical behavioral perspective of Gagné & Briggs (1974), Davis (1977, pp. 25–29) advanced eight criteria for the evaluation of teaching-learning experiences:

- the instructor establishes the context of the learning experience;
- appropriate technology is used to facilitate communication and aid student understanding and learning;
- the students’ active participation in the learning experience is elicited;
- the teacher provides different ways for the student to respond to the same concept, principle or idea;
- the teacher approaches concepts and principles from a variety of viewpoints;
- appropriate and specific feedback or knowledge of results is provided to students;
- the design of the learning experience is adapted to individual differences amongst students; and
- the subject matter is organized to enhance motivation and comprehension.

From an information-processing perspective, though centred in teachers’ transmission of information, Brown and Atkins (1988, pp. 8–16) analysed effective lecturing according to the following dimensions.

1. Intentions and planning: preparation, information of goals and work plans, etc.
2. Transmission:
 - clarity in explaining: organization (appropriate structuring and sequencing of topics, adequate quantity of content, etc), use of structuring moves (cues to signal the direction of the lecture, to emphasize key points and to link between topics – marking opening, closing and transitions of topics – and between the lecture and previous experience and knowledge), of schemata, organizational advances and final summaries, and of pause and emphasis (appropriate pace), establishment of internal and external relationships of the topic, etc.
 - expressiveness: this is the main way to promote students’ interest. It includes appropriate use of verbal and non-verbal communication (gestures, eye contact, voice inflection), of examples and analogies, and of the different types of explanation (narrative, anecdotal and conceptual), reflecting enthusiasm, dynamism, pleasantness, etc.
3. Receipt: keeping student’s attention, facilitating long-term memory (for instance, providing links to previous information).
4. Output: effective note-taking, students’ reactions, learning outcomes.

This exclusive focus on teacher behaviour is balanced by taking into account students’ approaches to learning (Marton & Säljö, 1976; Entwistle & Ramsden, 1983). From this perspective, Ramsden (1992, pp. 96–103) summarized the features of effective teaching in the following six principles:

- clarity of goals and intellectual challenge;
- quality of explanation and stimulation of student interest;
- promotion of independence, sense of control and active engagement of students in the learning process;
- concern and respect for the students and their learning (including teacher's versatility in teaching skills and his/her availability for student consults);
- learning from students (paying attention to their comments and analysing their learning outcomes); and
- appropriate feedback and evaluation.

Thus behavioral, informational and learning-centred models stress different aspects of teaching. Nevertheless, some basic aspects recur beyond distinctive theoretical bases, diversity of labelling and uneven degree of specification. These conceptual categories can be summarized in the five basic aspects of the quality of teaching performance in the classroom shown in Table 1.

Dimensions of teaching competence: applied research

The aforementioned convergence of conceptual categories as well as the dominance of 'knowledge transmission' orientation (*versus* 'learning facilitation' orientation, in terms of Kember & Gow, 1994) in higher education teaching, led to a marked overlap in the content of student questionnaires for teaching evaluation. Claims that typical rating forms are based upon transmissive models of instruction, and focus on conventional classroom teaching by teacher-centred methods, are frequently found in the literature (Centra, 1993; d'Apollonia & Abrami, 1997; McKeachie, 1997; Kolitch & Dean, 1999; Kember *et al.*, 2002). In spite of this homogeneity, the studies and reviews of student ratings of teaching performance in the English-speaking context provide contradictory results about the dimensionality of this construct. This outcome may result from specific contentuances, including the degree of specificity and length of the rating forms, from the different teaching situations to which they are applied and, finally, from the statistical technique used.

Table 1. Aspects of teaching competence derived from the theory of teaching-learning

-
1. Planning of teaching and preparation of lessons (the studies reviewed take subject mastery and actualization for granted), informing of intentions (goals, work plans, evaluation criteria) and appropriate workload (in quantity and difficulty).
 2. Communication skills and abilities required for the effective transmission of information: organization, clarity of explanation, stimulation of students' interest (expressiveness, use of examples and applications, diversity of approaches to the topic), students' involvement and active participation.
 3. Didactic and methodological resources, adequate to the subject and audience.
 4. Interaction with students: responsiveness to comments, availability for consults; attention to diversity of the class.
 5. Appropriate feedback for the orientation of students' learning.
-

Among the studies of multidimensional approaches, the work of Marsh (Marsh & Hocevar, 1984; Marsh, 1987, 1991; Marsh & Hocevar, 1991; Marsh & Dunkin, 1992; Marsh & Roche, 1997, among others) is outstanding because of its breadth and continuity over time. This work rests on a broad agreement among numerous researchers about the complexity of teaching and, especially, the need to reflect this feature in evaluations of formative purposes. In the aforementioned studies, Marsh identified a structure of nine principal factors which brought up four second-order factors (see Table 2).

This research could be criticized because of its focus on the statistical aspect, and specifically on factor analytical outcomes, to the detriment of conceptual analysis. The search for construct validity of students' ratings should be more theory-driven and based on more diverse analytical evidence.

In this regard, Feldman's (1988, 1989, 1996) research reflects a more conceptual approach. Feldman identified 22 basic dimensions which can be categorized – logically and empirically – in three dimensions reflecting the teacher roles: presenter, manager and facilitator.

The analysis of these multidimensional studies discloses a high degree of similarity in their instruments' items and in the first-order factors obtained. Likewise, these studies clearly suggest the feasibility of a general factor since, in many cases, the advanced second-order multifactor structure could be well-integrated into a single general factor. Thus, this empirical evidence may support the hypothesis of unidimensionality of these rating forms.

The authors that have perhaps most clearly shown the strong unidimensionality of students' rating forms are Abrami and d'Apollonia (Abrami, 1989; Abrami & d'Apollonia, 1990, 1999; Abrami *et al.*, 1996; d'Apollonia & Abrami, 1997). These authors pointed out that although teaching may be multidimensional, the perceptions of students reflected in the rating forms load on a global factor fairly strongly.

On more empirical grounds, they re-analysed factor analyses applied to diverse instruments by others. In their most systematic review (Abrami *et al.*, 1996), they found a very strong general factor (explaining 55% of the variance after rotation),

Table 2. Dimensional structure of Marsh's 'Student's Evaluation of Educational Quality' (SEEQ)

| <i>First-order factors</i> | <i>Second-order factors</i> |
|----------------------------|-----------------------------|
| Learning | I. Presentation |
| Teacher's enthusiasm | |
| Organization/Clarity | |
| Programme | II. Rapport |
| Interaction with group | |
| Individual interaction | III. Course material |
| Exams/Grades | |
| Assignments/Reading | III/IV |
| Workload/Difficulty | IV. Workload |

which comprises the role of the teacher as instructor, and another two factors (role of the teacher as a person and feedback) with less explanatory power and correlated with the first factor. It is interesting to note the similarity between these factors and the four second-order factors identified by Marsh, as well as their convergence with the conceptual synthesis advanced in the previous section. In sum, Abrami *et al.* (1996), though recognizing the multidimensionality of teaching competence, stress the powerful global dimension or component, and that the ratings on overall items would be appropriate for summative decisions.

These studies suggest the need for a pragmatic compromise, from a practical and operative position, conceptualizing the teaching construct as mainly unidimensional and, at the same time, comprising weaker specific dimensions. This position is already implicit in the works of both Marsh and Abrami and d'Apollonia, and is explicitly supported by McKeachie (1997). Furthermore, the stress on summative or formative purposes for teaching rating forms would give rise to different analysis and report needs, producing a single overall score or several specific scores respectively.

In a previous study in the Spanish context, Apodaca and Rodríguez (1999) compared the degree of fit of unidimensional and multidimensional solutions to the data from the rating form applied at the University of the Basque Country, by means of CFA. The multidimensional solution rendered a better goodness-of-fit, as well as operative advantages for the elaboration of feedback reports of results for formative purposes. This analysis identified five dimensions, labelled as 'Development of the programme', 'Mastery', 'Organization and clarity', 'Teacher-student interaction', 'Methodological-didactic resources' and 'Exams and grades' – which are clearly consistent with the dimensions advanced in the theoretical and applied studies reviewed above (Tables 1 and 2).

In a later work, Apodaca and Grad (2002) tested the compatibility of unidimensional and multidimensional solutions by comparing the structure of forms used at the University of the Basque Country (UPV) and the Autonomous University of Madrid (UAM), by means of parametric (CFA) and non-parametric (SSA) techniques of structural analysis. The results were not conclusive, though the structural analysis rendered interesting evidence in support of multidimensionality. The small number of items in the analysed instruments could constraint the chances of adequately verifying the multidimensional structure. Overcoming this limitation, the present work will compare the dimensional structure of large and short forms.

Parametric and non-parametric structural analysis

Parametric dimensional analysis

In the last ten years, there has been a huge increase in the use of the structural equations model (SEM – Batista and Coenders, 2000). Structural equations techniques distinguish between structural models and measurement models. This study applies the measurement model to carry out confirmatory factor analysis (CFA).

Besides the constraints of CFA on the variable distributions as a parametric technique, for the interpretation of our model's goodness-of-fit, it is worth recalling here

that the chi-square of a model tends, in general, to be more statistically significant (suggesting a poor fit for the data) the greater the sample size (Marsh *et al.*, 1988; Maruyama, 1998), the more restrictive and parsimonious the model and the greater the covariances between the variables. Thus, this technique paradoxically favours studies that use small samples, postulate complex models and are based on weak correlations (Apodaca and Páez, 1992). More specifically, as regards the present work, there are two aspects which simultaneously play as strengths and weaknesses of SEM technique: the manipulability of the fit of the measurement model and the hierarchical structure imposed on that model.

Firstly, regarding the model's manipulability, it is important to emphasize that SEM technique is only virtually confirmatory. This technique makes possible a large set of definitions of the model parameters, sometimes hardly sustained by the theory. Furthermore, in practice, the models are 'improved' by iterative testing. The obtained fit indexes play a fundamental role in these 'modifications' of the parameters. Thus, the model is modified, sometimes substantially, just by following the goodness-of-fit (Batista and Coenders, 2000). In this way, even though theory may have played an important role in the definition of the initial model, the model development is mainly data-driven. The extent of the observation of the theoretical framework depends on the researcher's judgement. All of this points clearly to an exploratory more than a confirmatory strategy. The confirmatory character does not arise so much from the technique itself as from the research design – for example, from strategies looking for model replication in the same or other samples (Maruyama, 1998).

Most of the flexibility and manipulability of SEM is due to the definition of the error parameters of the observed variables (items). Though these errors are usually defined as independent terms, some sources of error may simultaneously affect several items. Likewise, some factors or latent variables may affect certain items that are not part of the substance of the construct. These reasons may sometimes justify the specification of covariance structures between errors (Byrne, 1998).

However, in defining errors as intercorrelated, the unexplained variance of these models would be encompassed by these correlations. This strategy would lead to the paradox whereby the model may fit better because of these error intercorrelations than because of the predicted factors or latent variables.

Non-parametric structural analysis

Structure similarity analysis is a technique of ordinal multidimensional scaling developed by Guttman (Lingoes & Guttman, 1967; Guttman, 1968). This technique spatially represents any matrix of coefficients of association between the variables, so that the distances between the points are a monotonous function of these coefficients. Specifically, the variables are represented as points in a multidimensional space, in which the distances between the points are an inverse function of their empirical relationships (measured by any coefficient of association or distance). The greater the empirical association between two variables, the closer their locations in the spatial representation (for an introduction to SSA, see Canter, 1985; Borg & Lingoes, 1987;

Shye *et al.*, 1994; Borg & Groenen, 1997; for a comparison of SSA with other multi-dimensional scaling, factor and cluster analyses, see Dillon & Goldstein, 1984; Canter; 1985; Davison, 1985; Shye, 1988).

It is expected that the variables theoretically interrelated will converge empirically, allowing the partition of SSA space into discrete regions, and that the structural relationships will be reflected in the spatial configuration of these regions (Lingoes, 1981). Thus, these configurations of points and regions in space are interpreted to verify the theory (a 'configurational verification' approach, in Davison's (1983) terms). The significance of the result basically depends on the possibility of interpreting substantially the spatial representation by applying the *a priori* theory. The statistical validity of the outcomes is only assured by replication of findings.

SSA allows the identification of linear as well as hierarchical and circular relationships, with full, partial or no internal order (see Brown, 1985, for typical configurations of these spatial configurations).

Since the different aspects of teaching performance are qualitative categories of a similar conceptual level and, in general, without internal order (except for the possible practical conflict between concluding the programme and paying attention to the interaction with the students), we expect them to be organized in a polar configuration – that is, each aspect of teaching performance would emerge in a circular arrangement of wedge-like regions emanating from a common origin of a two-dimensional space in SSA.

Method

A sample of 7318 rating forms from the Universidad del País Vasco/Euskal Herriko Unibertsitatea (UPV) and another sample of 90,905 rating forms from the Universidad Autónoma de Madrid (UAM) were analysed. In both cases, students filled out a rating form for each teacher from whom they received classes.

The samples analysed represent different types of population. On the one hand, the sample from the UAM covers first- to fifth-year students from almost all the study areas (sciences, social sciences and humanities, except medicine) of a medium-sized general university in the Spanish context. Teaching methods ranged from traditional to participative lecturing (Marton & Säljö, 1976; Entwistle & Ramsden, 1983). On the other hand, the sample from UPV is smaller and homogeneous because it comes from a law faculty with conventional lecturing as the dominant teaching method.

In order to verify the generalizability of the latent structure of the studied constructs, the structures of data obtained with two different rating forms were compared. On the one hand, the data from UPV were gathered in 1990, applying a large students' rating form of teaching based directly on the work of Tejedor *et al.* (1987, 1988). The considerable length of the rating form applied at this time makes it especially appropriate for the goals of this study. The data from UAM, on the other hand, were gathered in 1999 using a shorter form.

Both questionnaires shared a focus on teacher performance in lecturing. Though this feature would disregard other aspects of the teaching-learning process, it was

consistent with the hegemony of a teacher-centred transmissive model of instruction (cf. Kember & Gow, 1994) and conventional lecturing methods in Spanish higher education, as well as with the original aim of this evaluation procedure to provide feedback for improving personal teaching skills.

The rating form applied at the UPV included 50 items. Sixteen items were dropped for these analyses, because they reflect the dimensions of fulfilment of teachers' formal duties (class attendance and punctuality) and exercises, as well as those items with a non-response rate higher than 10%. The overall rating items were also dropped from the analyses, as we considered that they would favour unidimensional solutions. Table 3 shows the items of the UPV rating form, organized by aspects of teaching performance.

Applying the same criteria, we analysed 13 of the 17 items in the UAM rating form, which does not include items reflecting the 'assessment' dimension. Table 4 presents the items organized by aspects of teaching performance.

The statistical analysis were carried out by means of the structural equations model (confirmatory factor analysis, CFA) of the AMOS software (Arbuckle & Wothke, 1997) and of similarity structure analysis (SSA – Guttman, 1968) non-parametric multidimensional scaling.

Results

Confirmatory Factor Analysis (CFA)

The structure of the rating forms from the UPV and the UAM was analysed on the basis of the conceptual model presented above. For this purpose, the items were classified in the five categories shown in Table 1. The analyses suggested the need to subdivide some of these broad categories, this subdivision being theoretically coherent as well as empirically defensible.

CFA of the UPV data. The results of the analysis of the sample of ratings form from UPV are presented in Figure 1. Estimations were made using the maximum likelihood (ML) algorithm. The estimated parameters of the factor loadings on each of the items, as well as the loadings of the general factor on the first-order factors, clearly show both distinct covariance structures for specific competencies and a strong general factor – common to all the instrument items. The goodness-of-fit indexes of the model are shown in Figure 1, and merit some clarifying remarks.

In general, we can consider the fit of the data to the model as moderate. Nevertheless, this judgement should be qualified, since the goodness-of-fit found cannot be evaluated in absolute terms, but rather in relation to data characteristics and alternative models. As regards the characteristics of the data, there are two important aspects to borne in mind. On the one hand, the large sample size implies that Chi-square based indicators of fit are likely to have an extremely low associated probability (Marsh *et al.*, 1988). On the other hand, the correlations among the items included in the analysis are fairly high, and this also negatively affects some of the fit indexes.

Table 3. Rating form applied at the UPV – teaching performance (34 items)

-
- 1. Planning and preparation*
- 1.1. Planning*
1. Teacher's programme covers the most important aspects of the subject
 2. The teacher informs the course programme
 4. (He/she) completes the full programme
 5. (He/she) distributes the programme uniformly throughout the academic period
 8. (He/she) cites literature that is usually current
 9. (He/she) seems to be au fait with recent developments in the subject field
- 1.2. Mastery*
7. He/she seems to have mastery of the subject taught
 27. (He/she) seems self-confident with respect to the topics explained
- 2. Communication skills and abilities*
- 2.1. Organization and clarity*
6. He/she takes care that concepts are understood
 10. (He/she) explains clearly
 11. (His/her) classes are well prepared
 12. (His/her) presentations have a well-defined scheme
 13. (He/she) responds with accuracy and precision to questions asked
 14. (He/she) summarizes the main points of his/her exposition
 15. (He/she) usually makes clear the aspects he/she considers important, taking time over those that are difficult to understand
 16. (He/she) Is a good public speaker
- 2.2. Enthusiasm and interest*
21. (He/she) gets the students interested in the subject
 25. His/her teaching style makes the subject interesting
 26. (He/she) seems to like teaching
 28. (He/she) takes care to give good classes
- 3. Interaction with students*
3. The teacher discusses the programme content and objectives with students
 17. (He/she) promotes discussion in class
 18. (He/she) achieves student participation in class
 19. (He/she) foments criticism of his/her own ideas
 20. (He/she) talks to students about how the class is going
 22. (He/she) takes into account the students' opinion when making decisions on issues related to the progress of the course
 23. (He/she) takes interest in his/her students
 24. (He/she) tends to help students who have difficulties
- 4. Didactic and methodological resources*
29. (He/she) proposes appropriate complementary study material (texts, notes, ...)
 30. (He/she) uses in class didactic material (schemata, graphs, ...) that helps to better understand the expositions
- 5. Assessment*
31. His/her exams are conceived so that one learns by taking them
 32. His/her exams link to what has been taught
 33. Grades obtained by his/her students fit with their knowledge
 34. The exam is designed basically to verify comprehension
-

Response scale: 1 – 'do not agree at all'; 2 – 'little agreement'; 3 – 'moderately agree'; 4 – 'fairly strongly agree'; and 5 – 'totally agree'

Table 4. Rating form applied at the UAM – teaching performance (13 items)

-
1. *Planning and preparation*
- 1.1. *Planning*
7. (He/she) has provided information about the programme and the work-plan for the course
8. In general, the teacher has followed the scheduled work-plan
9. (He/she) has provided information on the evaluation activities and criteria for the subject taught
- 1.2. *Preparation and mastery*
5. (His/her) classes are well-prepared
6. The teacher seems to have mastery of the subject and be au fait with progress in the field
2. *Communication skills and abilities*
- 2.1. *Organization and clarity*
1. The teacher explains clearly
3. (He/she) usually makes clear the aspects he/she considers important
- 2.2. *Enthusiasm and Interest*
2. (He/she) takes care to ensure that students learn
4. (He/she) contributes to making the subject interesting
3. *Interaction with students*
10. The teacher has a receptive attitude to students' questions or suggestions
11. (He/she) foments student participation in the class
4. *Didactic and methodological resources*
12. Theoretical concepts are adequately complemented with examples, text commentaries, exercises, problems, assignments, etc.
13. Bibliography and/or reading material indicated by the teacher are useful for studying the subject
-

Response scale: 1 – 'totally disagree'; 2 – 'fairly strongly disagree'; 3 – 'disagree'; 4 – 'neither agree nor disagree'; 5 – 'agree'; 6 – 'fairly strongly agree'; 7 – 'totally agree'

The fit indexes of the model were selected according to two criteria. First, they are the indexes most commonly used in this field. Second, they represent the main families of fit statistics – overall measures and parsimony adjusted measures of goodness-of-fit. Specifically, we report chi-square for hypothesis testing, root mean square of the model residual (RMR) and goodness of fit index (GFI) as overall absolute measures, and the parsimony goodness of fit index (PGFI) as measure adjusted for model complexity.

According to Byrne (1998), values of GFI higher than .90, PGFI higher than .50 and RMR lower than .05 would indicate acceptable fit, even in models with significant chi-square statistics. According to these criteria, the data analysed show appropriate (or close to appropriate) levels of fit. It is noteworthy that the PGFI in particular shows an especially high goodness-of-fit. This confirms that the combination of high levels of item covariance and large number of items reduces the values of the absolute measures of fit. In sum, given the power of the model, its relative parsimony and the large number of indicators employed, the obtained goodness-of-fit seems particularly high.

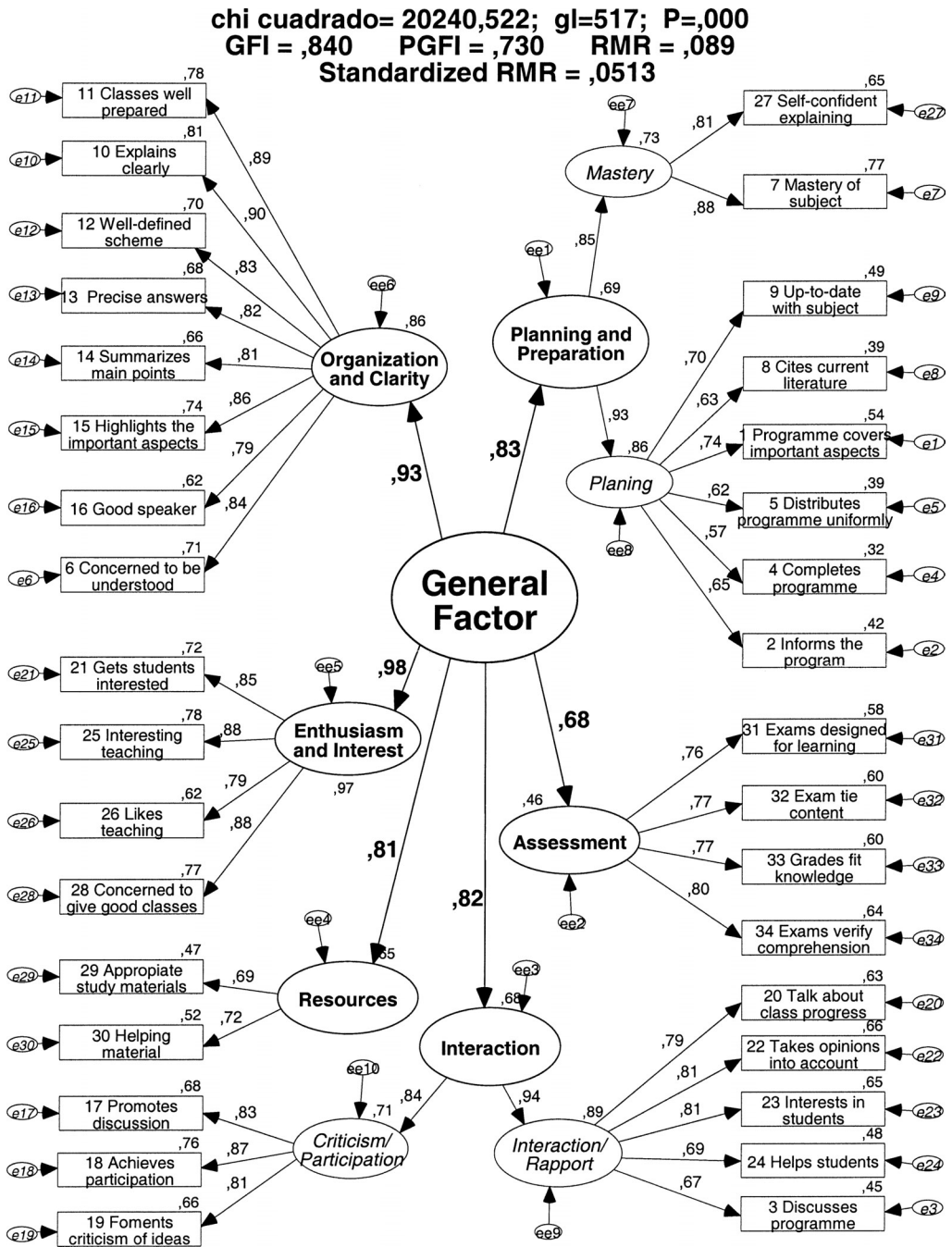


Figure 1. CFA of the UPV rating form – teaching performance (34 items): standardized results

The results are clarifying with respect to the dimensionality of the latent structure. As Figure 1 shows, dimensions 1, 2 and 4 of the conceptual scheme presented in Table 1 are split into more specific aspects. Thus, the aspect of 'planning' is distinguished from 'mastery' within the category 'planning and preparation'. The category 'communication skills and abilities' is split into two aspects with strong internal consistency: 'organization and clarity' and 'enthusiasm and interest'. General aspects ('criticism/participation') are distinguishable from interpersonal aspects ('interaction/rapport') of the dimension 'interaction with students'. Finally, the category 'didactic and methodological resources' has limited reliability, as it is represented by just two items.

Almost all the item loadings in these first-order factors are higher than .70, supporting the internal consistency of these specific dimensions. At the same time, for the split categories, the first-order factors show very high loadings on their corresponding second-order factors – showing that the split does not challenge the conceptual consistency of the dimension initially proposed. The first-order factors 'organization and clarity' and 'enthusiasm and interest' emerge as directly associated with the general third-order factor.

The general factor shows very high loadings on these two first-order factors (.93 and .98) whilst presenting high but comparatively lower loadings on the rest of the factors. This association may indicate the relative centrality of these dimensions for the construct of the students' perception of teaching performance. This asymmetry in the generality or centrality of the different dimensions raises doubts over whether the hierarchical structure tested in the CFA is the most appropriate for reflecting the latent structure of teaching performance.

CFA of the UAM data. Verification of this latent structure in the UAM data is difficult due to the smaller number of items in the applied rating form. The structure of the data is similar to that advanced in Table 1 – maintaining the distinction between 'preparation' and 'planning', but not between 'enthusiasm and interest' and 'organization and clarity' (see the corresponding CFA in Figure 2).

The results show moderate but appropriate goodness-of-fit according to the criteria of Byrne (1998). Comparing UAM and UPV models, it comes out that the PGFI of the UAM solution is lower, while the rest of the statistics (GFI, standardized RMR) are notably better than in the UPV solution. This difference supports the hypothesis that a large number of indicators per dimension can significantly reduce the unweighted statistics, and that the fit indexes should be assessed jointly and taking into account the specific features of the advanced model.

The resulting structure for the UAM data is quite similar to that for UPV, showing clearly a multidimensionality compatible with a very strong general factor. In this case also the dimension 'communication skills and abilities' receives the greatest loading on the general factor. Similarly, in general, the specific dimensions show different degrees of centrality. This asymmetry in centrality may be compatible with structures distinct from the hierarchical ones.

Chi-square = 28380. 688; df=60; P=.000
GFI = .952 PGFI = .628 RMR = .056
Standardized RMR = .0289

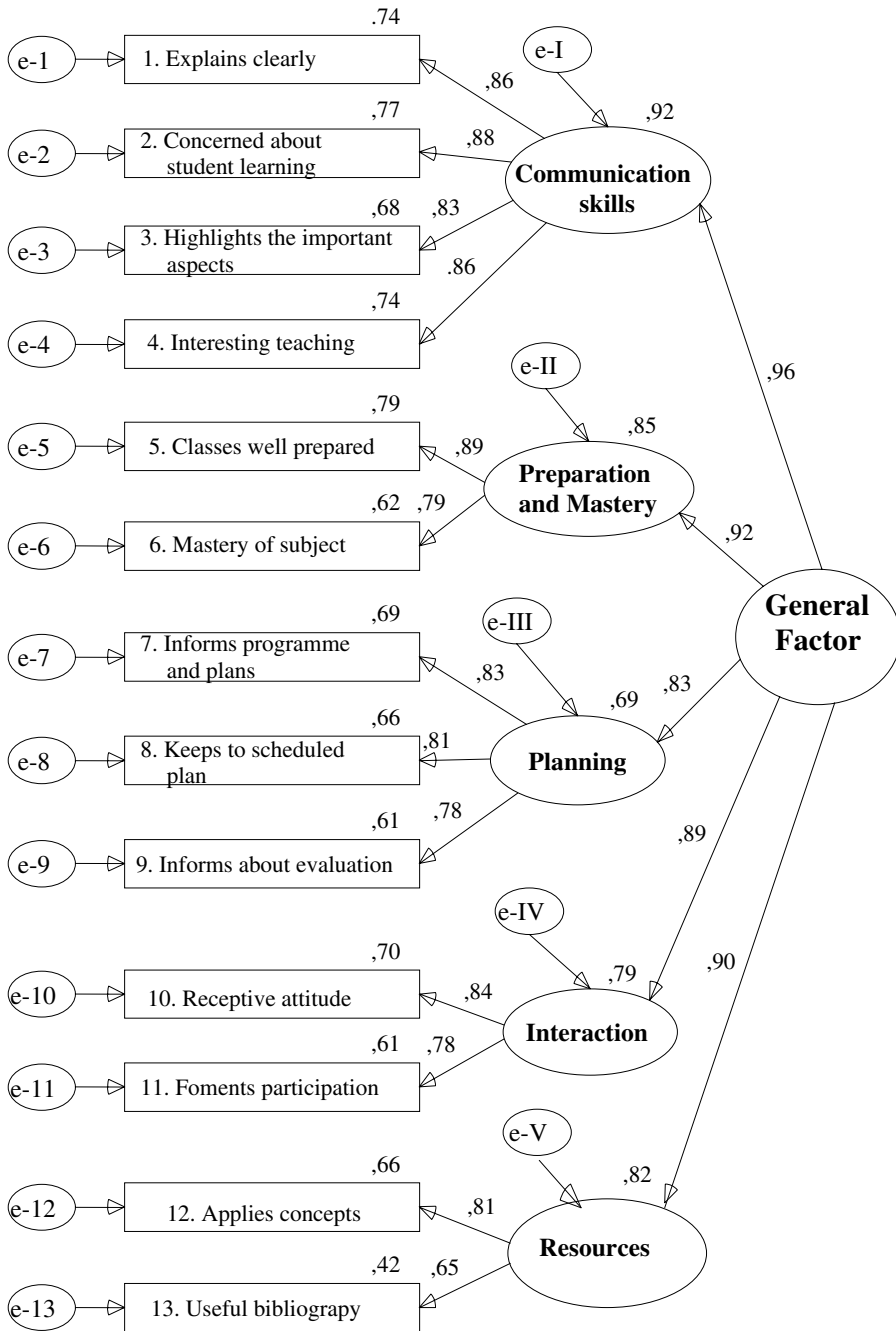


Figure 2. CFA of the UAM rating form – teaching performance (13 items): standardized results

The convergence of the findings between the forms applied at UPV and UAM is especially remarkable due to the smaller number of items on the UAM form. This form presents a stricter selection of items, and this, in turn, constrains the coverage of certain dimensions (such as 'assessment') and the representativeness of the indicators of each specific first-order dimension.

Multidimensional scaling: Similarity Structure Analysis (SSA)

In order to verify the polar configuration of the aspects of teaching performance, the same data were subjected to a two-dimensional SSA. The fit of the spatial representation was acceptable in both samples, due to alienation coefficients of .19 for the UPV and of .14 for the UAM data. Both representations could have been interpreted conceptually by identifying the different aspects of teaching performance reflected in these rating forms. In both samples, the items reflecting the same aspect of teaching performance converge in the space, allowing us to distinguish discrete circular sectors (Figures 3 and 4).

In line with expectations, it was possible to define regions (in this technique, only the points are produced by the analysis: the lines and borders are subsequently determined by the researcher) with a polar configuration in both representations. The items for the different aspects of teaching emerge in distinct sectors, which intersect close to the centre and open up towards the periphery of the representation. The absence of overlapping between the conceptually defined regions supports the internal consistency of the advanced categories. It is possible to distinguish specific sectors for those dimensions that have been identified in the CFA for UPV. That is the case for 'preparation and mastery' and 'planning' as aspects of 'planning and preparation', and for 'organization and clarity' and 'enthusiasm and interest' as aspects of 'communication skills and abilities' (both in the SSA of the UPV and the UAM data). In the region of 'interaction with students' of the UPV we could, likewise, have distinguished between items reflecting 'criticism/participation' (items 17, 18, 19) and more specific items reflecting 'rapport' (items 20, 22, 23, 24). This spatial configuration of students' ratings is consistent with the multidimensional models of teaching competence.

Moreover, the polar configurations found are partially ordered by the practical contradiction between 'interaction with students' (rapport, fomenting participation, attention to opinions and suggestions) and completion of the programme (quantity of information transmitted) in the students' view. The analyses of both samples show the more instrumental items referring to fulfilment of the programme (such as item 4, 'completes the full programme' at UPV, and item 8, '... has followed the scheduled work-plan' at UAM) are situated in opposition to the more expressive items referring to the teacher's personal interaction (such as item 23, 'takes an interest in his/her students' and 24, 'tends to help students who have difficulties' at UPV and item 11, 'fomenting student participation in class' at UAM). Furthermore, although the UAM rating form does not include 'assessment' items, its SSA shows an empty space in the position of the circular order (between the regions of 'planning' and 'resources') where this aspect emerges in the SSA of UPV. Consequently, the order of the circular

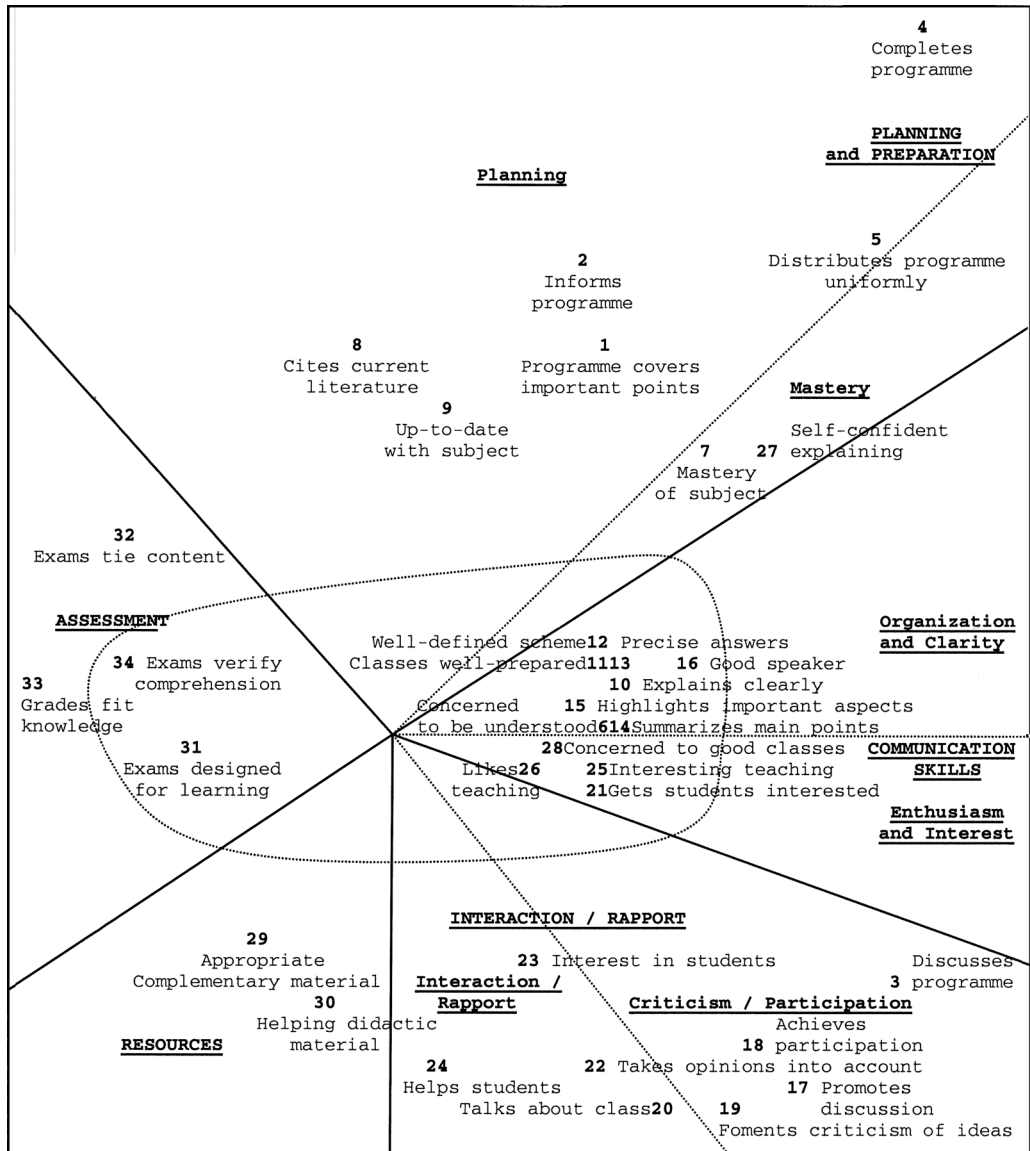


Figure 3. SSA of the UPV rating form – teaching performance (34 items): 2 dimensions – alienation coefficient: .19

sectors (that is, of the aspects of teaching performance) can be considered identical in the two samples.

The SSAs of both samples also point out the different centrality of the various aspects of teaching competence. Thus, the items of the ‘organization and clarity’ and ‘enthusiasm and interest’ dimensions emerge close to the centre, whilst the items of the remaining aspects emerge more towards the periphery of the representations. This asymmetry suggests the possibility that teaching competence has a

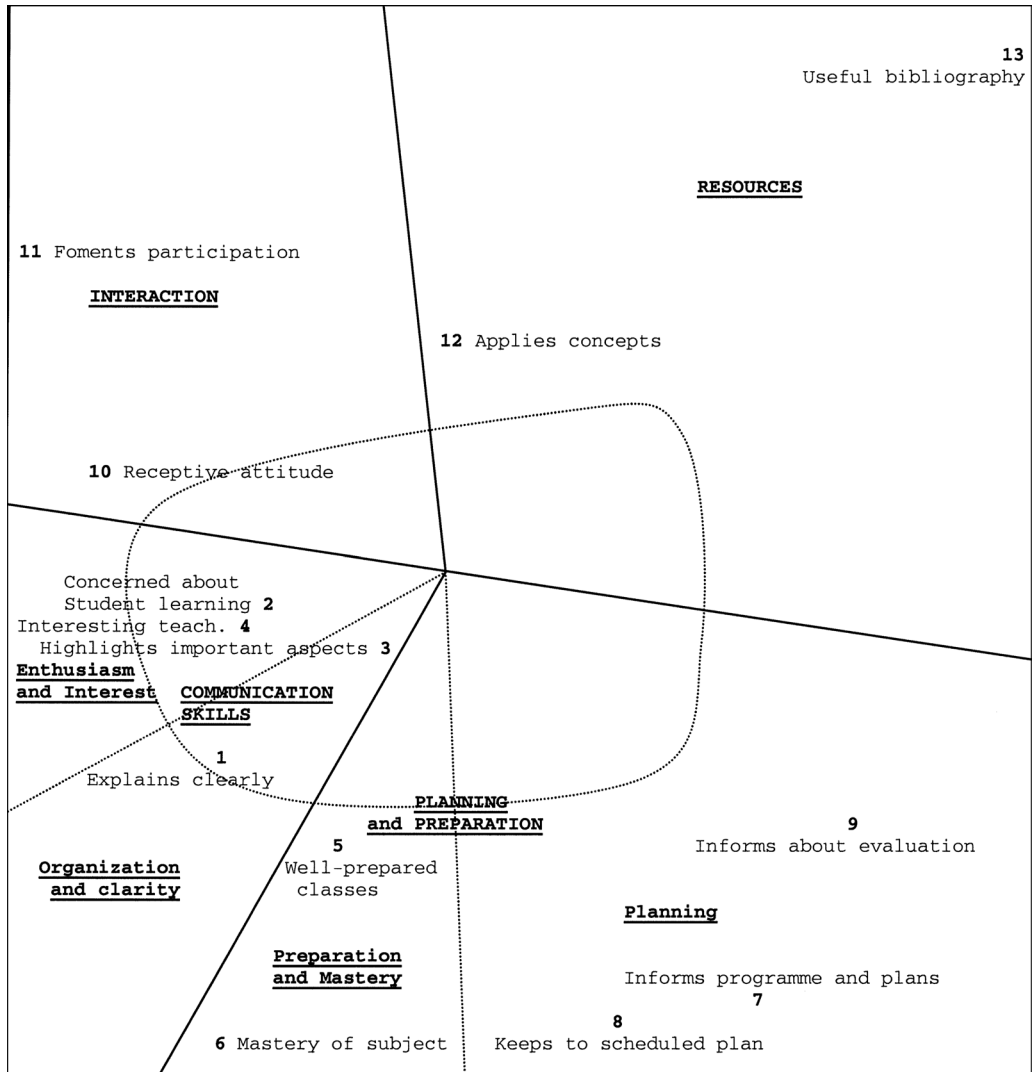


Figure 4. SSA of the UAM ratings form – teaching performance (13 items): 2 dimensions – alienation coefficient: .14

radex configuration, combining a polar and a modular (centre-periphery) facet in terms of facet theory (Brown, 1985). The items' content suggests that the relative centrality may reflect their contribution to the quality of teaching performance, according to the students' view. Thus, the most central items would contribute more, while the more peripheral items would play a more subsidiary role, in teaching effectiveness.

In sum, there clearly emerges a polar (that is, multidimensional in factorial terms), partially ordered, configuration and a marked asymmetry in the centrality of the

different specific dimensions for the construct of teaching competence. The aspects of 'enthusiasm and interest' and 'organization and clarity' would be central for this construct, in line with the results of the CFA that indicated considerable empirical overlapping between these dimensions and the general factor. The rest of the specific dimensions would have a more peripheral role, and be more weakly related to teaching competence as perceived by students.

In conclusion, integrating the outcomes of CFA and SSA in the two samples and rating forms allows us to verify that:

- students' ratings have a multidimensional structure compatible with a very strong general factor
- some of the specific dimensions ('enthusiasm and interest', 'organization and clarity') show a high level of generality or centrality, so that they are hardly distinguished from the general factor from an empirical point of view
- the hierarchical structure may not be the most appropriate for these instruments' data, due to the asymmetry in the generality or centrality of the specific dimensions.

Discussion

The outcomes of CFA and SSA have disclosed that the scales of teaching competence show multidimensionality compatible with a strong general factor. Both analyses have uncovered meaningful differences in the relationship of the specific dimensions to the general factor of the construct. SSA seems to be more effective for reflecting these structural differences, because of its superiority for representing the non-hierarchical structure of the construct and the different relevance or centrality of its indicators. In this context, the results of the SSAs point to a radex structure – resulting from the combination of a partially-ordered polar facet (the aspects of teaching competence) and another, modular facet (the centrality of each aspect for the construct). Thus, SSA emerges as a good heuristic for the comprehension of teaching competence and for the generation and testing of theories on its structure.

The replication, in both samples, of the circular order of the teaching competence aspects in the polar facet – contrasting (quantity of) information transmitted against interaction with students – suggests some relations of conflict and compatibility among the aspects of teaching competence in the students' view. This structural regularity merits additional studies both from conceptual and empirical perspectives.

Our interpretation of the modular facet suggests that the relative centrality of the different aspects of the teaching competence construct would reflect their contribution to learning according to students' perception. On the one hand, some specific dimensions (such as 'enthusiasm and interest' and 'organization and clarity') emerge as more central in the SSA, and achieve almost complete correlations and are empirically confounded with the general factor in the CFA. The structural centrality and empirical overlap of these dimensions with the general factor are consistent with the nuclear role attributed to these aspects by the theory on effective teaching. Thus, these empirical and conceptual arguments would support a unidimensional model of teaching

competence. On the other hand, the rest of the teaching competence dimensions (such as 'planning and preparation', 'interaction with students', and 'didactic and methodological resources') would have a more peripheral or specific role. The distinction of these dimensions in the current analyses, however, would support the empirical and conceptual multidimensionality of the construct of teaching performance.

The general convergence of the findings, despite the differences in the instruments and the populations studied, supports the existence of a common latent structure for teaching competence. However, the structural nuances found suggest the need for further research on this issue. First of all, it would be convenient to replicate these results with forms designed for verifying systematically the radex configuration of teaching competence. This selection of items would also require previous studies for assessing, conceptually and empirically, the contribution of the different aspects of teaching performance to students' learning. For this purpose it seems to us especially necessary to carry out studies on the predictive validity of students' ratings in multi-section designs, using the students' achievement as the criterion. In this way, it would also be possible to estimate the predictive validity of each specific dimension of teaching competence included in these rating forms. Finally, these studies should control and vary more systematically the characteristics of the teaching situation and student population.

Although the results of this study with respect to the centrality or periphery of the different aspects of teaching performance are based exclusively on students' opinions, we would like to underscore that the structure of these perceptions is consistent with the theory of effective teaching. Thus, we can conclude that this evidence supports the construct validity of these opinions. Nevertheless, the limitations of students as expert observers of all the aspects of teaching advise against using these opinions as the only source of evaluative information on teaching competence.

There are clear implications of this new path of analysis for the summative and formative functions of the evaluation processes for teaching quality. Let us point out some of the most important ones:

With regard to the formative function, the results endorse the use of specific scores for each one of the teaching competence dimensions, but distinguishing between the more central and the more peripheral or subsidiary dimensions. Indeed, the theory and experience of teacher training show that the aspects of organization, clarity and stimulation of students' interest – just like the technical competencies related to management and planning – can certainly be improved with appropriate counselling and training (e.g., Marsh & Roche, 1993; Stevens & Aleamoni, 1985; Piccinin *et al.*, 1999). On the other hand, competencies more related to the teacher's attitudinal, cultural and personal characteristics – such as enthusiasm or interaction/rapport – are especially resistant to change, and this makes intervention policies in these domains difficult.

Experience shows, however, that interventions focused on the aspects of organization and planning have potential contradictory effects when teaching staff perceive that the competencies required and rewarded by the institution are those more related to conventional lecturing – to the detriment of other, more innovative competencies,

such as those related to the interaction/rapport aspect. On the positive side, the greater generality and resistance to contextual effects of the 'enthusiasm' and 'organization and clarity' dimensions make them a more suitable focus of intervention to improve teaching quality.

We might also suggest some important implications of this study for the summative function of teaching quality evaluation by means of students' ratings. Up to now, the debate was polarized between two positions defending, on the one hand, the use of scores in each specific dimension, and on the other, the use of a single global score that could be constructed on the basis of overall rating items. This study supports both this strategy and the more sophisticated alternative of constructing overall scores by weighting properly the scores of the specific dimensions – according to their centrality for the teaching competence construct. Indeed, the convergence between the centrality of the 'enthusiasm' and 'organization and clarity' aspects (and their overlap with the general factor of the teaching competence construct), and the feasibility of their improvement through training, recommends the inclusion of specific scores of these dimensions in the reports of teaching quality evaluation – both because of the importance of this specific feedback and because of the stimulation of feasible improvements in teaching performance.

In any case, the design of initial and in-service training programmes for university teaching staff should focus on personal orientations and competencies. These programmes would be more costly and have longer duration, but they could have a far greater impact on teaching quality than the mere learning of some didactic techniques.

Acknowledgement

This research was supported by project grant BSO2000-1407 of the National Plan of R&D (2000–2003) of the Spanish Ministry of Education and Science.

References

- Abrami, P. C. (1989) How should we use student ratings to evaluate teaching? *Research in Higher Education*, 30, 221–227.
- Abrami, P. C. & d'Apollonia, S. (1990) The dimensionality of ratings and their use in personnel decisions, in: M. Theall & J. Franklin (Eds) *Student ratings of instruction: issues for improving practice* (San Francisco, CA, Jossey-Bass), 97–112.
- Abrami, P. C. & d'Apollonia, S. (1999) Current concerns are past concerns, *American Psychologist*, 54, 519–520.
- Abrami, P. C., d'Apollonia, S. & Rosenfield, S. (1996) The dimensionality of student ratings of instruction: what we know and what we do not, in: J. C. Smart (Ed.) *Higher education: handbook of theory and research* (Vol. XI) (New York, Agathon Press), 213–264.
- Aparicio, J. J., Tejedor, J. & Sanmartin, R. (1982) *La enseñanza universitaria vista por los alumnos: Un estudio para la evaluación de los cursos en la enseñanza superior* (Madrid, Instituto de Ciencias de la Educación de la Universidad Autónoma de Madrid).
- Apodaca, P. & Grad, H. (2002) Análisis dimensional de las opiniones de los alumnos universitarios sobre sus profesores: comparación entre técnicas paramétricas y no-paramétricas, *Revista de Investigación Educativa*, 20, 385–410.

- Apodaca, P. & Páez, D. (1992) Los modelos estructurales, posibilidad y limitaciones, in: D. Páez, J. Valencia, J. F. Morales, B. Sarabia & N. Ursúa (Eds) *Teoría y método en psicología social* (Madrid, Anthropos).
- Apodaca, P. & Rodríguez, M. (1999) La opinión de los alumnos en la evaluación de la calidad docente: posibilidades, limitaciones y estructura dimensional de sus indicadores, in: J. Vidal (Coord.) *Indicadores en la universidad: información y decisiones* (Madrid, Consejo de universidades, Ministerio de Educación y Cultura).
- Arbuckle, J. L. & Wothke, W. (1999) *Amos user's guide: versión 4.0* (Chicago, IL, SPSS).
- Batista, J. M. & Coenders, G. (2000) *Modelos de ecuaciones estructurales* (Madrid, La Muralla).
- Borg, I., & Groenen, P. J. F. (1997) *Modern multidimensional scaling: theory and applications* (New York, Springer).
- Borg, I., & Lingoes, J. (1987) *Multidimensional similarity structure analysis* (New York, Springer-Verlag).
- Borg, I. & Shye, S. (1995) *Facet theory: form and content* (Thousand Oaks, CA, Sage).
- Brown, G. & Atkins, M. (1988) *Effective teaching in higher education* (London, Routledge).
- Brown, J. (1985) An introduction to the uses of facet theory, in: D. Canter (Ed.) *Facet theory: approaches to social research* (New York, Springer-Verlag), 17–57.
- Byrne, B. M. (1998) *Structural equation modeling with LISREL, PRELIS and SIMPLIS: basics concepts, applications, and programming* (London, Lawrence Erlbaum Associates).
- Canter, D. (Ed.) (1985) *Facet theory: approaches to social research* (New York, Springer-Verlag).
- Centra, J. (1993) *Reflective faculty evaluation* (San Francisco, CA, Jossey-Bass).
- Cohen, P. A. (1981) Student ratings of instruction and student achievement: a meta-analysis of multisection validity studies, *Review of Educational Research*, 51, 281–309.
- d'Apollonia, S. & Abrami, P. (1997) Navigating student ratings of instruction, *American Psychologist*, 52, 1198–1208.
- Dancer, L. S. (1990) Introduction to facet theory and its applications, *Applied Psychology: An International Review*, 39, 365–377.
- Davis, R. H. (1977) Learning by design, in: J. A. Centra (Ed.) *Renewing and evaluating teaching* (San Francisco, CA, Jossey-Bass).
- Davison, H. (1983) *Multidimensional scaling* (New York, Wiley).
- Davison, M. L. (1985) Multidimensional scaling versus component analysis of test intercorrelations, *Psychological Bulletin*, 97, 94–105.
- de Miguel, M. (1991) *Criterios para la evaluación del profesorado universitario* (Oviedo, Departamento de Ciencias de la Educación de la Universidad de Oviedo).
- de Miguel, M. (1998) La evaluación del profesorado universitario. Criterios y propuestas para mejorar la función docente, *Revista de Educación*, 315, 67–84.
- de Miguel, M. (2003) Calidad de la enseñanza universitaria y desarrollo profesional del profesorado, *Revista de Educación*, 331, 13–34.
- Dillon, W. R. & Goldstein, M. (1984) *Multivariate analysis: methods and approaches* (New York, Wiley).
- Elejabarrieta, F. & Perera, S. (1989) Las paradojas de los rigores metodológicos en psicología social, *Revista Vasca de Psicología*, 2(1), 81–90.
- Entwistle, N. J. (1987) *Understanding classroom learning* (London, Hodder and Stoughton).
- Entwistle, N. J. & Ramsden, P. (1983) *Understanding student learning* (London, Croom Helm).
- Feldman, K. A. (1976) The superior college teacher from the student's view, *Research in Higher Education*, 5, 243–288.
- Feldman, K. A. (1988) Effective college teaching from the student's and faculty's view: matched or mismatched priorities? *Research in Higher Education*, 28, 291–344.
- Feldman, K. A. (1989) Instructional effectiveness of college teachers themselves, current and former students, colleagues, administrators, and external (neutral) observers, *Research in Higher Education*, 30, 137–193.

- Feldman, K. A. (1996) Identifying exemplary teaching: using data from course and teacher evaluations, *New Directions for Teaching and Learning*, 65, 41–50.
- Gagné, R. M. & Briggs, L. J. (1974) *Principles of instructional design* (New York, Holt, Rinehart and Winston).
- Garcia Ramos, J. M. (1997) Análisis Factorial Confirmatorio en la validación del constructo competencia docente del profesor universitario, *Bordón*, 49(4), 361–391.
- Gillespie, K. H., Hilsen, L. R. & Wadsworth, E. C. (Eds) (2002) *A guide to faculty development: practical advice, examples, and resources* (Bolton, MA, Anker Publishing Company).
- Guttman, L. (1968) A general nonmetric technique for finding the smallest coordinate space for a configuration of points, *Psychometrika*, 33, 469–506.
- Guttman, L. (1982a) Facet theory, smallest space analysis, and factor analysis, *Perceptual and Motor Skills*, 54, 491–493.
- Guttman, L. (1982b) 'What is not what' in theory construction, in: R. M. Hauser, D. Mechanic & A. Haller (Eds) *Social structure and behavior* (New York, Academic Press), 331–348.
- Guttman, R. & Greenbaum, C. W. (1998) Facet theory: its development and current status, *European Psychologist*, 3, 13–36.
- Kember, D. & Gow, L. (1994) Orientations to teaching and their effect on the quality of student learning, *Journal of Higher Education*, 65, 59–74.
- Kember, D., Leung, D. and Kwan, K. (2002) Does the use of student feedback questionnaires improve the overall quality of teaching? *Assessment and Evaluation in Higher Education*, 27, 411–425.
- Kolitch, E. & Dean, A. V. (1999) Student ratings of instruction in the USA: hidden assumptions and missing conceptions about 'good' teaching, *Studies in Higher Education*, 24, 27–42.
- Lingoes, J. C. (1981) Testing regional hypotheses in multidimensional scaling, in: I. Borg (Ed.) *Multidimensional data representations: when and why* (Ann Arbor, MI, Mathesis), 280–310.
- Lingoes, J. C. & Guttman, L. (1967) Nonmetric factor analysis: a rank-reducing alternative to linear factor analysis. *Multivariate Behavioral Research*, 2, 485–505.
- Lowman, J. (1984). *Mastering the Techniques of Teaching*. (San Francisco: Jossey-Bass).
- Marsh, H. W. (1987) Students' evaluations of university teaching: research findings, methodological issues, and directions for future research, *International Journal of Educational Research*, 11, 253–388.
- Marsh, H. W. (1991) Multidimensional student's evaluation of teaching effectiveness: a test of alternative higher-order structures, *Journal of Educational Psychology*, 83, 285–296.
- Marsh, H. W. & Dunkin, M. J. (1992) Students' evaluations of university teaching: a multidimensional perspective, in: *Higher education: handbook of theory and research* (Vol. VIII) (New York, Agathon Press), 143–233.
- Marsh, H. W. & Hocevar, D. (1984) The factorial invariance of student evaluations of college teaching, *American Educational Research Journal*, 21, 341–366.
- Marsh, H. W. & Hocevar, D. (1991) The multidimensionality of students' evaluation of teaching effectiveness: the generality of factor structures across academic discipline, instructor level, and course level, *Teaching & Teacher Education*, 7, 9–18.
- Marsh, H. W. & Roche, L. A. (1993) The use of students' evaluations and an individually structured intervention to enhance university teaching effectiveness, *American Educational Research Journal*, 30, 217–251.
- Marsh, H. W. & Roche, L. A. (1997) Making students' evaluations of teaching effectiveness effective, *American Psychologist*, 52, 1187–1197.
- Marsh, H. W., Balla, J. R. & McDonald, R. P. (1988) Goodness-of-fit indexes in confirmatory factor analysis: the effect of sample size, *Psychological Bulletin*, 103, 391–410.
- Marton, F. & Säljö, R. (1976) On qualitative differences in learning, outcome, and process, *British Journal of Educational Psychology*, 46, 4–11.
- Maruyama, G. M. (1998) *Basics of structural equation modeling* (Thousand Oaks, CA, Sage).
- McKeachie, W. J. (1997) Student ratings: its validity of use, *American Psychologist*, 52, 1218–1225.

- Newble, D. & Cannon, R. (1995) *A handbook for teachers in universities & colleges: a guide to improving teaching methods* (London, Kogan Page).
- Piccinin, S., Cristi, C. & McCoy, M. (1999) The impact of individual consultation on students' ratings of teaching, *International Journal for Academic Development*, 4, 75–88.
- Quintanilla, M. A. (1998) En pos de la calidad: notas sobre una nueva frontera para el sistema universitario español, *Revista de Educación*, 315, 85–95.
- Ramsden, P. (1992) *Learning to teach in higher education* (London, Routledge).
- Rodríguez, S. (2003) Nuevos retos y enfoques en la formación del profesorado universitario, *Revista de Educación*, 331, 67–100.
- Seldin, P. (1995) *Improving college teaching* (Bolton, MA, Anker Publishing Company).
- Shye, S. (1988) Inductive and deductive reasoning: a structural reanalysis of ability tests, *Journal of Applied Psychology*, 73, 308–311.
- Shye, S., & Elizur, D., with Hoffman, M. (1994) *Introduction to facet theory: content design and intrinsic data analysis in behavioral research* (Thousand Oaks, CA, Sage).
- Stevens, J. J. & Aleamoni, L. M. (1985) The use of evaluative feedback for instructional improvement: a longitudinal perspective, *Instructional Science*, 13, 285–304.
- Tejedor, F. J. (1993) *Experiencias españolas de evaluación de la enseñanza universitaria y nuevas perspectivas*. III Jornadas nacionales de Didáctica Universitaria. Evaluación y desarrollo profesional. Las Palmas, Gran Canaria, Universidad de Las Palmas de Gran Canaria, Servicio de Publicaciones.
- Tejedor, F. J., Jato, E. & Minguez, C. (1987) Evaluación del profesorado universitario por los alumnos en la universidad de Santiago, in: M. J. Carrera Gonzalo (Eds) *Primeras jornadas nacionales sobre evaluación y mejora de la docencia universitaria* (Valencia, Universidad de Valencia), 133–210.
- Tejedor, F. J., Jato, E. & Minguez, C. (1988) Evaluación del profesorado universitario por los alumnos de la universidad de Santiago, *Studia Paedagogica*, 20, 73–134.