TYPE Original Research
PUBLISHED 29 August 2023
DOI 10.3389/feduc.2023.1229129

Check for updates

OPEN ACCESS

EDITED BY
Gavin T. L. Brown,
The University of Auckland, New Zealand

REVIEWED BY
Diana Pereira,
University of Minho, Portugal
Michalis P. Michaelides,
University of Cyprus, Cyprus

*CORRESPONDENCE
Juan Carlos Pérez-Morán
✉ jucarpint@gmail.com

RECEIVED 25 May 2023
ACCEPTED 11 August 2023
PUBLISHED 29 August 2023

CITATION
Henriquez PS, Pérez-Morán JC, del Cid García CJ and Zamora JE (2023) Factor structure and invariance of the scale to measure teaching performance in the area of social sciences.
Front. Educ. 8:1229129.
doi: 10.3389/feduc.2023.1229129

COPYRIGHT
© 2023 Henríquez, Pérez-Morán, del Cid García and Zamora. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Factor structure and invariance of the scale to measure teaching performance in the area of social sciences

Patricio Sebastián Henríquez[1], Juan Carlos Pérez-Morán[2]*, Carlos Javier del Cid García[1] and Jesús Enrique Zamora[1]

[1]Autonomous University of Baja California, Ensenada, Mexico, [2]Promoter Network of Diagnostic Evaluation Methods and Educational Innovation, Ensenada, Mexico

The use of scales to evaluate teaching from the students' perspective is a method frequently used in educational systems around the world. The objective of this study is to analyze the factorial structure of the Teaching Performance Evaluation Scale (EEDDocente, by acronyms in Spanish) designed with the purpose of providing information that favors decision-making based on evidence for the improvement of teaching in the area of Social Sciences, as well as measuring the invariance by *School stage* and *Educational Program*. The sample consisted of 1,849 students of the Bachelor's Degrees in Law, Psychology, Accounting, Administration, Education Sciences, Communication Sciences, Computer Science, and Sociology of the School of Social and Administrative Sciences (FCAyS) of the Autonomous University of Baja California, Mexico. Based on a three-factor model that meets the fit and quality criteria, a Multi-group Confirmatory Factor Analysis (MGCFA) was performed to measure the invariance of the EEDDocente by *School stage* and *Educational program*. It is concluded that the three-factor model can be used to measure, from the students' perspective, the performance of teachers in the *Area of Social Sciences*. Likewise, it is concluded that the invariance of the simultaneous measurement is achieved, providing evidence to perform mean difference analysis between the different *Educational programs*.

KEYWORDS

students' evaluation of teaching, higher education students, validity, confirmatory factor analysis, invariance

1. Introduction

The evaluation of teaching in Higher Education Institutions (HEIs) is one of the most relevant components linked to the assumption of improving educational quality (Calatayud, 2021; Torquemada, 2022; Bleiberg, 2023; Silva, 2023). The measurement of the effectiveness of teaching practice occupies a central place in HEIs strategies, which allows the generation of information on the teaching and learning process that serves as an input to trace routes for improving the quality, relevance, effectiveness and accountability of education systems around the world (Chen and Hoshower, 2003; Liebowitz, 2021; Seivane and Brenlla, 2021; Camacho, 2022; Zhao et al., 2022). In addition, it is a transcendental input for the improvement and feedback of teacher performance in its multiple dimensions, thus attending to the formative function of this process (Marsh, 2007; Luna and Torquemada, 2008; Liebowitz, 2022; Silva et al., 2022).

Around the world, accountability and the growing demand to ensure the improvement of learning of future professionals graduating from universities has placed the evaluation of teaching performance as the dominant axis of educational policies (Vaillant, 2016; Liebowitz, 2022). However, there is a growing interest in the methodological aspects, techniques and instruments for collecting information (questionnaires, attitude scales, interviews, focus groups, classroom observation), and the subjects (students, teachers, managers, external experts) best suited to obtain reliable, valid, sufficient and relevant data on the evaluation of teaching (Marsh, 1984; Cruz Ávila, 2007; Romero and Martínez, 2017; Zamora, 2021; Bleiberg, 2023).

In particular, there is growing concern about the use of Students' Evaluations of Teaching (SET) to make high-impact or consequential decisions in processes such as promotion, tenure, and awarding of stimuli and incentives (Boring et al., 2016; Hornstein, 2017; Wang and Guan, 2017; Benton, 2018; Ching, 2018; Mitchell and Martin, 2018; Bazán-Ramírez et al., 2021). Likewise, the purposes of the evaluation of teacher performance have been mainly oriented to condition the hiring or dismissal processes of academic staff, deciding who gets an economic incentive or job promotion based on the result of the teacher evaluation (Stroebe, 2016; Gómez and Valdés, 2019). However, more and more decision makers, education systems and HEIs see an opportunity and advantage in using relatively brief SETs, mid- or end-of-course, to provide formative feedback on teacher performance and competencies (Marsh, 2007; Silva et al., 2022).

The study of university teacher performance began to become widespread internationally in the 1980s, as part of the accountability processes derived from changes in government policies for financing higher education (Cisneros-Cohernour and Stake, 2010; Zamora, 2021; García-Olalla et al., 2022). The evaluation of teaching performance has its genesis in the first student learning assessment systems in the United States during the 1920s; by the second half of this decade, learning assessment served as a tool to evaluate teaching (Alcaraz-Salarirche, 2015; Zhao et al., 2022). For its part, the SET was an innovation in HEIs in the United States, a consequence of the consumer orientation of the capitalist system: students, as users of the educational service, are the ones who should evaluate it (García, 2000). During the 1980s and early 1990s, numerous studies were carried out on the subject, for example, those of Cohen (1981, 1983), Feldman (1988, 1989a,b, 1990, 1992, 1993), and Marsh (1984, 1986, 1987, 1993, 2007).

Teacher performance evaluation must maintain high and solid technical quality standards to fulfill its main purpose, which is linked to improving teaching practices and student learning. However, in the mid-1990s, studies began to emerge that questioned the reliability and usefulness of quantitative instruments to evaluate teaching (Theall et al., 2001; García, 2014; Boring et al., 2016; Benton, 2018; Ching, 2018). Among the most recurrent criticisms by researchers are those related to the idea that the SET presents problems of logic and structure in terms of components and characteristics to test the effectiveness of teaching, and random and biased responses, as well as a subjective judgment on the part of students about teaching (Stroebe, 2016, 2020; Wang and Guan, 2017; Ching, 2018; Zhou and Qin, 2018; Gu et al., 2021).

Despite criticisms to the contrary, it is undeniable that the use of scales and questionnaires has been the most widely used mechanism to evaluate university teachers (Wang and Guan, 2017; Zamora, 2021), and that questionnaires as evaluation instruments are viable tools to measure the effectiveness of teaching in HEIs (García, 2014; Mohammadi, 2021). However, it is necessary that evaluation questionnaires maintain validity, reliability, relevance, and pertinence for uses and consequences in the educational context (Messick, 1995; Kane, 2006; International Test Commission, 2013; Spooren et al., 2013; American Educational Research Association, 2018; Reyes et al., 2020; Lera et al., 2021).

The evaluation of the quality of teaching practice through experience, certifications, academic degrees, among other factors, shows little correlation with the effectiveness of teaching performance (Williams and Hebert, 2020). Thus, the evaluation of teaching, based on the perception of students, currently has a preponderant role in the processes of improving the quality of teaching in universities (Aleamoni, 1999; Salazar, 2008; Mohammadi, 2021; Zamora, 2021). The SET allows HEIs to have a reference for the improvement of teaching practice, as long as the performance measures maintain a high level of objectivity, methodological rigor and relationship with the implementation dimensions of the academic objectives of educational systems (House, 1998; Navarro and Ramírez, 2018; Seivane and Brenlla, 2021). At this point, it is important to mention that most of the criteria or dimensions of the SET are defined by committees of specialists, which are based on models of indicators of teaching quality and effectiveness, but with a strong influence of philosophical and pedagogical principles, and of the policy and regulations of the functions of the academic staff of each HEIs.

Among the first syntheses of criteria to design SET are those proposed by Feldman (1976) and Hildebrand et al. (1971). By analyzing students' points of view, Feldman (1976) proposed three categories for effective teaching: (a) *Presentation*, which includes teachers' enthusiasm for teaching, their knowledge of the subject, and clarity of presentation and organization of the course; (b) *Facilitation*, which refers to the availability of teachers for consultation, respect for students and their ability to encourage students to achieve learning in class; and (c) *Regulation*, which includes teachers' ability to set clear objectives, classroom management skills, appropriateness of course materials and activities, and fairness in student assessment and feedback. For their part, Hildebrand et al. (1971) and Hildebrand (1973) propose five factors to measure teaching effectiveness: (a) *Analysis and synthesis skills*, which refers to the teacher's mastery of class content; (b) *Clarity and organization*, which refer to the teacher's ability to present course topics; (c) *Interaction with the group*, which refers to the teacher's ability to interact with students and maintain the active participation of the group; (d) *Interaction with each student*, which refers to the teacher's ability to establish trust and respect with each individual student; and (e) *Dynamism and enthusiasm*, which refers to the teacher's enthusiasm and pleasure in teaching the subject. More recently, authors such as Marsh (1987), Marsh and Dunkin (1997), Richardson (2005), and Schellhase (2010) have proposed models of up to nine to 10 criteria (*assignments and readings, breadth of coverage, examinations and grading, group interaction, individual rapport, instructor enthusiasm, learning and academic value, organization and clarity, workload and difficulty,* and *summative evaluation*) to assess the quality of instruction.

In Ibero-America, several authors have focused on the design and validity of the measurement of teacher performance through scales considering various criteria models of the effectiveness and quality of teaching; in particular, on obtaining the psychometric properties of the measurement instruments, and on the evidence of their internal consistency and reliability. In this sense, the studies by García-Gómez-Heras et al. (2017), in Spain; Estrada et al. (2019), in Nicaragua; Bazán-Ramírez et al. (2021), in Peru; and Márquez and Madueño (2016), and Bazán-Ramírez and Velarde-Corrales (2021), in Mexico; they are noteworthy. For their part, García-Gómez-Heras et al. (2017) focused on revealing which professors' behaviors were most appreciated by first-year students of studies taught at the Faculty of Health Sciences of the Rey Juan Carlos University of Madrid (Degrees in Medicine, Nursing, Physiotherapy, Dentistry Psychology and Occupational Therapy). The authors applied the questionnaire developed by Tuncel (2009) on the behaviors of university teachers that influence the academic performance of students. This questionnaire is made up of 48 items organized into six factors: (a) *Emotional aptitude of university teachers*, (b) *Teacher-student interaction*, (c) Achievement of educational objectives, (d) *Theory-practice relationship*, (e) *Organization and planning*, and (f) *Feedback*.

Likewise, Estrada et al. (2019), Gómez and Valdés (2019) conducted a study to establish the psychometric properties of the *Opinion Questionnaire for the Evaluation of Teaching Performance* (OQETP) composed of 38 items, focused on evaluating teaching performance from the students' perception, at the National University of Trujillo, Nicaragua. The OQETP items are presented on a Likert-type scale with five response categories and are organized into eight questionnaire dimensions: (a) *Formal Responsibility*, (b) *Methodology*, (c) *Communication*, (d) *Materials*, (e) *Attitude*, (f) *Evaluation*, (g) *Motivation*, and (h) *Satisfaction*.

In Peru, Bazán-Ramírez et al. (2021) analyzed the factorial structure of the *Teaching Performance Scale for Psychology Teachers* (EDDPsic) and measured the invariance between groups (according to *gender*, *age* and *academic stage*). This instrument was designed based on the model of five didactic performance criteria (Carpio et al., 1998; Silva et al., 2014). In total, the EDDpsic is made up of 18 items ($K = 18$) organized into subscales: (a) *Competence Exploration* ($k = 3$), (b) *Criteria Explanation* ($k = 5$), (c) Illustration ($k = 3$), (d) Feedback ($k = 4$), and (e) Evaluation ($k = 3$). Their study involved 316 Psychology students, from basic cycles (fourth and sixth semesters) and disciplinary-prof*essional cycles* (eighth and tenth semesters), from two public universities in Peru. They also performed a Multigroup Confirmatory Factor Analysis (MGCFA) with the five-factor model that showed the best fit indices. Based on their results, they determined the invariance of the scale measure across the three study variables (*age, sex* and *academic stage*), for which the participants were divided into independent groups. The results revealed adequate fitness tests for the *Configural* model in each of the three variables (χ^2 $p > 0.05$, CFI < 0.01, RMSEA ≤0.06), so it was considered that the structure of the model was the same for each group. Similarly, evidence of factorial invariance was obtained for the *Weak* (M1), *Strong* (M2) and *Strict* (M3) models, in the variables of *age* (M1: CFI = −0.004; M2: CFI = -0.001; M3: CFI = −0.001), and *gender* (M1: CFI = −0.001; M2: CFI = −0.001; M3: CFI = −0.001). In the case of the *academic stage* variable, evidence of invariance was obtained for the *Weak* and *Strong* models (M1: CFI = −0.004; M2: CFI = −0.000) but not for the *Strict* model (M3: CFI = −0.018).

In Mexico, Márquez and Madueño (2016) analyzed the psychometric properties of an instrument made up of 16 items ($K = 16$) applied to students at a university in Sonora, to recover their opinion on the basic competencies of teachers in the teaching of undergraduate courses. From the 30,224 questionnaires answered, the construct validity of the instrument was determined using the principal components method with Varimax rotation, extracting two factors: (a) *Pedagogical mediation* ($k = 11$), and (b) *Teaching attitudes* ($k = 5$). For their part, Bazán-Ramírez and Velarde-Corrales (2021) evaluated the performance of teachers and students in their didactic interactions through the self-report of 124 psychology students in Mexico. The authors obtained the construct validity, convergent and divergent, of five *didactic performance criteria*, both of the teacher and the student, by means of EFA and CFA. The validation confirmed the theoretical structuring of five factors that correspond to the five dimensions: (a) *Exploration of competencies,* (b) *Explicitness of criteria,* (c) *Illustration,* (d) *Feedback*, and (e) *Evaluation*, derived from the models of didactic performance (Carpio et al., 1998; Silva et al., 2014). The authors also performed descriptive analyses of the students' responses to the didactic performance criteria, according to their *academic stage*, *age* and *sex*.

In summary, it can be said that the models of criteria and instruments to evaluate teaching in the HEIs present a wide diversity of components and characteristics. Likewise, these instruments generally present acceptable psychometric properties of reliability and validity. However, most of them are made up of a large number of criteria and items, which results in instruments that can help in a diagnosis with greater depth and granulation, but make it difficult to apply in cases where students have to answer repeatedly an instrument for each of their teachers at the end of each school year and throughout their university studies. Although, it is important to highlight that most measurement models based on more than five criteria do not satisfactorily meet all the necessary technical quality criteria. In this regard, several authors mention that one of the problems of SET is that multidimensional models that try to cover a large number of criteria based on theory present internal structure problems (Stroebe, 2016, 2020; Ching, 2018). This is explained to some extent when universities include in their teacher evaluations criteria that refer to affective components such as *student satisfaction with the class*, *interest in the subject content*, *teacher's capacity for empathy*, among others. So far it can be concluded that the instruments for measuring the effectiveness and quality of teaching that seek to include a wide variety of criteria present problems of logic and internal structure, as well as difficulties for their application in evaluation strategies where it is required that students respond repeatedly at a specific time in the school year. Another important problem to mention is that most of the SETs evaluate different criteria, making it impossible to make comparative studies that help to evaluate the policies to improve the quality of teaching among different educational programs, schools, and universities.

This paper analyzes the psychometric properties and evidence of construct validity of internal structure and invariance of the Teaching Performance Evaluation Scale (EEDDocente, by acronyms in Spanish) applied at the middle of each school stage to assess the performance of each one of the teachers in the different educational programs of the School of Administrative and Social Sciences (FCAyS, for its acronym in Spanish) of the Autonomous University of Baja California (UABC, for its acronym in Spanish). The EEDDocente is applied biannually with the purpose of identifying strengths and weaknesses of teaching performance from the students' perspective and thus provide feedback on teaching and design teacher training and updating courses (Cashin and Downey, 1992; Liebowitz, 2021; Zamora, 2021; Silva et al., 2022).

Despite the variety of instruments for the evaluation of teaching practice, the relevance of the EEDDocente lies in its purpose, design and objective that seek to maintain coherence between the instrument and the use of results (Stroebe, 2016; Estrada et al., 2019; Gómez and Valdés, 2019; Aravena-Gaete and Gairín, 2021). The EEDDocente was designed to provide information to identify teachers' needs for updating and continuous training, and to influence the improvement of performance and teaching practices at the classroom level. Among its specific characteristics, the EEDDocente focuses on student-centered teaching and, based on this, the information provided by the scale seeks to generate processes of reflection on teaching practice and a change in the conception and vision of how they develop university teaching (Tomás-Folchy and Durán-Bellonch, 2017).

However, there is no evidence related to the internal structure and invariance of this instrument. This paper aims to address this problem and contribute to the existing literature by analyzing the internal structure of the three-factor model of a reduced version ($K=15$) of the EEDDocente that is based on categories of solid theoretical models: (a) *Classroom organization*, (b) *Teaching quality*, and (c) *Learning assessment/feedback* (Hildebrand et al., 1971; Feldman, 1976; König et al., 2017; Nasser-Abu, 2017; Chan, 2018; Bazán-Ramírez et al., 2021; Henríquez et al., 2023). Likewise, with invariance analysis it is possible to reduce student bias when evaluating teaching among the different *educational programs* in which they are enrolled.

2. Method

2.1. Participants

We analyzed the responses of a focused sample of 1849 students out of a total of 4,226 enrolled in the FCAyS of the UABC who participated in the internal strategy of teaching performance evaluation 2022–1 (conducted in the first semester of the year). For the selection of the sample of participants, the FCAyS Teaching Evaluation Coordination randomly selects, during the second semester of each *school stage*, two groups from each semester of the eight current educational programs (Law, Psychology, Accounting, Business Administration, Educational Sciences, Communication Sciences, Computer Science, and Sociology), one from the morning shift and one from the afternoon shift. In addition, it randomly chooses four groups of the *Common core of the Areas of knowledge of Legal Sciences, Accounting and Administrative Sciences* and *Social Sciences*, two from the morning shift and two from the afternoon shift. Table 1 shows the distribution of the study sample by *Educational program, Common core* and *Area of knowledge*. Note that the number of participants by subject area shows a wide difference. In particular, between the *Area of Legal Sciences*, with 366 participating students, and the other two *Areas of knowledge*, where almost twice as many students participated in each of them. Likewise, there is a considerable difference between the sample of participating students per *School stage* [*Basic stage* (1st and 2nd semester) $N=632$, *Disciplinary stage* (3rd, 4th, 5th and 6th semester) $N=816$, *Terminal stage* (7th and 8th semester) $N=392$].

2.2. Measurement

Scale for the Evaluation of Teaching Performance (EEDDocente) was designed by the coordinators of teacher evaluation of the FCAyS (Henríquez et al., 2017, 2018; Henríquez and Arámburo, 2021) with the purpose of providing at the middle of each school stage relevant information, based on the opinion of the students, on the performance of each teacher who teaches classes in the current educational programs, favoring continuous training and decision-making to improve teaching. In total, a student can answer the EEDDocente up to seven times, depending on the number of teachers who teach the different classes in the semester in progress. The EEDDocente is a typical performance test made up of 25 ordered response items ($K=25$) with four categories: 1 = *Strongly disagree*, 2 = *Disagree*, 3 = *Agree*, 4 = *Strongly agree*. During the design of the EEDDocente, a committee made up of teachers and graduates of the Social Sciences area of the University where the scale is applied was formed, who participated together with specialists in the writing of the items. With this, it was sought to ensure that the scale items were designed from a student-centered teaching approach. The items are organized into three subscales in which the dimensions underlie: (a) *Course organization*, refers to the teacher's ability to explain in a clear and organized manner the contents of the subject matter and the objectives and activities in class, as well as to use didactic strategies in an adequate manner to awaken the students' interest in the learning objectives; (b) *Quality of teaching*, refers to the teacher's ability to relate the contents of the subject matter with those of other classes,

TABLE 1 Distribution of the sample of participating students by *Educational program, Common core* and *Area of knowledge* of the FCAyS.

Area of knowledge	Educational program	Population (%)	Sample (%)
Accounting and Administrative Sciences $N=1,404$ (33.2%)	Bachelor's Degree in Business Administration	496 (11.7%)	293 (15.8%)
	Bachelor's Degree in Accounting	390 (9.2%)	295 (16%)
	Bachelor's Degree in computer science	92 (2.2%)	57 (3.1%)
	Common trunk of the area of accounting and administrative sciences	426 (10.1%)	162 (8.8%)
Legal Sciences $N=1,174$ (27.8%)	Bachelor's Degree in Law	1,174 (27.8%)	366 (19.8%)
Social Sciences $N=1,648$ (39%)	Bachelor's Degree in communication Sciences	191 (4.5%)	117 (6.3%)
	Bachelor's Degree in Education Sciences	255 (6%)	111 (6%)
	Bachelor's Degree in Psychology	696 (16.5%)	287 (15.5%)
	Bachelor's Degree in Sociology	34 (0.8%)	26 (1.4%)
	Common trunk of the area of social sciences	472 (11.2%)	135 (7.3%)
FCAyS	Total students	4,226 (100%)	1849 (100%)

TABLE 2 Items of the Scale for the Evaluation of Teaching Performance (EEDDocente).

Factores	Items ID	Items of the EEDDocente
		The teacher...
F1. Course organization	Q4.1	explains the contents of the subject clearly and presents them with an appropriate sequence. [explica los contenidos de la materia con claridad y los presenta con una secuencia adecuada.]
	Q4.2	explain at the beginning of each class the objective and activities of the day. [explica al inicio de cada clase el objetivo y actividades del día.]
	Q4.10	has theoretical command of the subject. [tiene dominio teórico de la asignatura.]
	Q4.11	use digital tools appropriately. [utiliza las herramientas digitales de manera adecuada.]
	Q4.12	maintains timely communication with students. [mantiene comunicación oportuna con los estudiantes.]
	Q4.13	demonstrates communication skills and ease of speech. [demuestra habilidades de comunicación y facilidad de palabra.]
	Q4.14	respect the plan and the framework of the course. [respeta el plan y el encuadre del curso.]
	Q4.15	resolves student doubts appropriately. [resuelve dudas de los estudiantes de manera pertinente.]
	Q4.16	reflects commitment and enthusiasm for their work. [refleja compromiso y entusiasmo por sus labores.]
F2. Quality of Teaching	Q4.3	promotes the connection of the contents of the subject with situations, experiences or problems of daily life (for example, situations to be faced in the work context). [promueve la conexión de los contenidos de la materia con situaciones, experiencias o problemas de la vida cotidiana (por ejemplo, situaciones a enfrentar en el contexto laboral).]
	Q4.4	relates the contents of the subject with the contents of other subjects. [relaciona los contenidos de la materia con los contenidos de otras materias.]
	Q4.5	encourages student participation in class development, for example, through questions, presentations, discussion of ideas, opinions, etc. [fomenta la participación de los estudiantes en el desarrollo de la clase, por ejemplo, a través de preguntas, exposiciones, debate de ideas, opiniones, etcétera.]
	Q4.6	establishes rules and norms of socialization with students. [establece reglas y normas de socialización con los estudiantes.]
	Q4.7	fosters an atmosphere of coexistence based on trust and respect among all. [fomenta un ambiente de convivencia basado en la confianza y respeto entre todos.]
	Q4.8	make adaptations at the request or in favor of the learning needs of the group. [realiza adecuaciones a petición o a favor de las necesidades de aprendizaje del grupo.]
	Q4.9	awakens the group 's interest in the contents and purposes of the subject. [despierta el interés del grupo por los contenidos y propósitos de la asignatura.]
F3. Evaluation and Feedback of Learning	Q10.1	use evaluation strategies that I like. [utiliza estrategias de evaluación que me agradan.]
	Q10.2	use evaluation methods with which I learn better. [utiliza métodos de evaluaciones con los que aprendo de mejor forma.]
	Q10.3	promote forms of learning support parallel to partial exams: for example, advice, clarifications, doubts, post-evaluation feedback, among others. [promueve formas de apoyo al aprendizaje paralelas a los exámenes parciales: por ejemplo, asesorías, aclaraciones, dudas, retroalimentación post-evaluación, entre otras.]
	Q10.4	is interested in improving student learning, beyond the final grade obtained. [se interesa en el mejoramiento del aprendizaje de los estudiantes, más allá de la calificación final obtenida.]
	Q10.5	is concerned with establishing forms of evaluation related to real-life problems. [se preocupa por establecer formas de evaluación relacionadas con problemáticas de la vida real.]
	Q10.6	is concerned with differentiating between students who learn more and less easily, adapting their teaching strategies and forms of evaluation. [se preocupa por diferenciar entre los estudiantes que aprenden con mayor y menor facilidad, adaptando sus estrategias de enseñanza y las formas de evaluación.]
	Q10.9	shows openness for corrections and adaptations with respect to non-conforming grades or errors. [muestra apertura para correcciones y adecuaciones respecto con calificaciones inconformes o errores.]

Source: Self-made.

encourage group participation in class activities, establish norms of coexistence in the classroom and make adjustments to favor the achievement of the group's learning objectives; and (c) *Evaluation and feedback of learning,* refers to the teacher's ability to apply strategies for evaluation and feedback of learning with a formative approach, differentiate between students who learn more and less easily, adapt their teaching strategies and forms of evaluation, establish forms of evaluation related to real-life problems, and show openness to corrections and adjustments regarding non-conforming grades or errors. As a foundation, the EEDDocente is based on multidimensional conceptual models consolidated and commonly reported in the literature related to the evaluation of teaching by students (Marsh, 1984, 1993, 2007; Feldman, 1988, 1993; Centra, 1993; Braskamp and Ory, 1994; Arreola, 2007; Fink, 2008; Bazán-Ramírez et al., 2021). Table 2 shows the items that make up the three subscales of the EEDDocente.

2.3. Procedure

The protocol and procedure for applying the instrument was approved by the FCAyS-UABC Management and supervised by the FCAyS Teacher Evaluation Coordination, in accordance with current institutional research ethical standards. It should be noted that the application of the EEDDocente is part of the internal strategy of evaluation of the teaching performance of the FCAyS that is applied at the middle of each school cycle by the Coordination of Teaching Evaluation of said faculty. In particular, the application is carried out with the support of students who provide their professional social service and who are previously trained to apply the evaluation instrument in the classroom. The students to whom the instruments are applied are previously informed of the objectives and procedures of the evaluation strategy, and of the confidentiality, safeguarding, and use of their answers in order to promote the continuous training of teachers, research, and decision-making to improve the performance of teachers who teach classes at the FCAyS. On this occasion, the EEDDocente was administered during school hours in each of the classrooms of the 80 randomly selected groups that make up the sample. On average, the explanation of the purpose of the teacher evaluation, the instructions and the application of the EEDDocente lasted 25 min. In addition, at the end of the application an effort was made for the students to answer all the items on the scale.

2.4. Data analysis

The data analysis is organized in four stages: (1) purification of the database, descriptive statistics, elimination of atypical cases; (2) verification of the preliminary assumptions of normality, reliability and linearity; (3) explained variance, measure of sample adequacy and analysis of the internal structure through the application of the Confirmatory Factor Analysis (CFA); and (4) measurement of invariance using Multi-Group Confirmatory Factor Analysis (MGCFA). Following the recommendations of Hu and Bentler (1999) and Hirschfeld and Von-Brachel (2014), statistical analyses were performed with the support of the *dplyr* (Wickham et al., 2019), *psych* (Revelle and Revelle, 2015), *lavaan* (Rosseel, 2012) and *semTools* (Jorgensen et al., 2022) from the open source software RStudio version 1.4 (R Core Team, 2022).

In the first stage, the database was cleaned, eliminating missing and atypical cases based on the Tukey Fences test. As a result of said procedure, 1,679 of the 1,849 original cases remained, of which 549 are from the *Basic stage*, 748 from the *Disciplinary stage*, and 374 from the *Terminal stage*. Consecutively, the mean, standard deviation, standard error and *item-total* correlation (*rpbis*) of each one of the items, and the *general index* and by subscales of the EEDDocente were estimated.

In the second stage, the assumption of normality was tested by applying the *Multivariate normality* test for asymmetry and kurtosis by Mardia (1970) with an acceptance criterion ≥0.05. The Kolmogorov–Smirnov test with Lilliefors correction was performed consecutively. With the kurtosis coefficient it is possible to identify the tendency of the participants to respond in a biased way toward one of the response categories (Vance et al., 1983), while with the symmetry coefficient the degree of concentration of responses to a central area of the distribution. In the Kolmogorov–Smirnov test, if the value of p is less than α (0.05, default value), the null hypothesis is rejected (the distribution is normal) (Dallal and Wilkinson, 1986). As a result of said procedure, items Q4.3, Q4.7, Q4.10, Q4.12, Q4.13, Q4.14, Q4.15, and Q4.16 were eliminated, which presented values well outside the boundaries of the kurtosis and skewness coefficients between −1 and + 1 recommended by Hair et al. (2019). Likewise, items Q10.7 and Q10.8, which did not meet the cutoff criterion of $rpbis \geq 0.2$, were eliminated (Brown, 2012).

For its part, global and subscale reliability was verified with the estimation of the standardized *Rho Alpha* coefficient (ρ) and the McDonald *Omega* coefficient (ω) together with Cronbach's *Alpha* (α) (Cronbach, 1951, 1988; McNeish, 2018; Raykov and Marcoulides, 2019). The quality criteria for the reliability coefficients were $\rho \geq 0.70$, $\omega \geq 0.80$, and $\alpha \geq 0.70$ (Cronbach, 1951, 1988; Katz, 2006; Zhang and Yuan, 2016; Nájera-Catalán, 2019). Once the preliminary analysis and the quality criteria were taken into account, we proceeded to analyze the model of three factors [*Course Organization* (F1), *Quality of Teaching* (F2), and *Evaluation and Feedback of Learning* (F3)] that underlie in the internal structure of the EEDDocente through the CFA application. For this, the Weighted Least Squares (WLS) and Robust Weighted Least Squares (WLSMV) estimation methods were applied (Jöreskog and Sörbom, 1979; Brown, 2015; Kline, 2015; Gazeloglu and Greenacre, 2020). On the other hand, in the evaluation of the adjustment indexes, the recommendations of Hu and Bentler (1999) and Hair et al. (2019). In particular, the adjustment indices and criteria were the Comparative Fit Index (CFI) ≥ 0.95, the Tucker-Lewis Index (TLI) ≥ 0.95, the Normalized Mean Square Residual (SRMR) ≤ 0.08 and the Mean Square Error of Approximation (RMSEA) ≤ 0.06 (Browne and Cudeck, 1993; Schreiber et al., 2006). For the subsequent analysis, only items with factor loadings ≥0.43 were considered.

Finally, an MGCFA was carried out to measure the invariance by *School stage* and *Educational program* based on the adjusted model of three factors of the EEDDocente. To verify the assumption of invariance depending on the *School stage*, three groups were considered: (a) students of the *Basic stage* (1st and 2nd semester), (b) students of the *Disciplinary stage* (3rd, 4th, 5th and 6th semester), and students of the *Terminal stage* (7th and 8th semester). On the other hand, to verify the assumption of invariance based on the *Knowledge Area*, three groups were considered: (a) students enrolled in the *Accounting and Administrative Sciences Educational* programs, (b) students enrolled in the *Legal Sciences Educational* programs, and (c) students enrolled in the *Educational* programs of *Social Sciences*. The *Configural*, *Weak*, *Strong* and *Strict* invariance models were contrasted (Dimitrov, 2010; Milfont and Fischer, 2010). For this, the recommendations of Byrne et al. (1989) and Vandenberg and Lance (2000) focus the analysis on the increasingly restrictive comparison of the model parameters. To consider the factorial invariance between models adequate, it was established as a criterion that the Chi-square difference ($\Delta\chi^2$) was not significant ($p > 0.05$). However, since the $\Delta\chi^2$ is affected by the sample size, the recommendations of Vandenberg and Lance (2000), Cheung and Rensvold (2002) and Dimitrov (2010) were followed, establishing RMSEA parameters close to the cutoff criterion of 0.08, a difference in RMSEA parameters between models less than 0.015 (ΔRMSEA ≤0.015), and a difference in CFI and TLI parameters between models less than 0.010 (ΔCFI and ΔTLI <0.010) (Chen, 2007; Dimitrov, 2010; Putnick and Bornstein, 2016).

3. Results

3.1. Descriptive results and preliminary analyses

The average of the general index of the EEDDocente was 86.61, with a standard deviation of 11.05. Likewise, the average score of

TABLE 3 Descriptive statistics ($n = 1,679$, $K = 15$).

Subscale	Item (k)	M	SD	Skewness	Kurtosis	*rpbis*
1. Course organization	Q4.1	3.56	0.69	−1.34	0.64	0.69
	Q4.2	3.31	0.87	−0.98	−0.08	0.63
	Q4.9	3.44	0.78	−1.23	0.62	0.75
	Q4.11	3.62	0.69	−1.80	2.57	0.46
2. Quality of Teaching	Q4.4	3.07	0.95	−0.58	−0.83	0.53
	Q4.5	3.57	0.72	−1.61	1.79	0.55
	Q4.6	3.40	0.84	−1.23	0.55	0.58
	Q4.8	3.46	0.80	−1.34	0.90	0.69
3. Evaluation and Feedback of Learning	Q10.1	3.43	0.68	−1.09	1.04	0.75
	Q10.2	3.29	0.74	−0.83	0.27	0.76
	Q10.3	3.38	0.73	−1.03	0.67	0.75
	Q10.4	3.49	0.68	−1.18	0.88	0.73
	Q10.5	3.49	0.66	−1.09	0.77	0.69
	Q10.6	3.26	0.82	−0.89	0.09	0.70
	Q10.9	3.39	0.73	−1.07	0.80	0.60
Average		3.41	0.76	−1.15	0.8	0.66
General index of the EEDDocente		86.61	11.05	−1.10	1.02	

TABLE 4 Overall and subscale internal consistency values of EEDDocente.

Subscale	α	ρ	ω
1	0.77	0.77	0.78
2	0.75	0.75	0.77
3	0.90	0.90	0.90
Overall	0.92	0.92	0.93

the scale items was 3.41 (4 = *Strongly agree*), with item Q4.4 being the one with the lowest average score (3.07) and Q4.11 the one with the highest average score. (3.62). For its part, the average *item-total* correlation of the scale was 0.64, meeting the cut-off criterion (*rpbis* ≥ 0.2). Likewise, the items presented, on average, moderate correlations among themselves (0.42) with correlation coefficients that oscillated between 0.26 and 0.74. Table 3 shows the descriptive results of the items and the general index of the EEDDocente.

Regarding the assumption of normality, significant results ($p < 0.001$) were obtained with the multivariate normality test of asymmetry and kurtosis by Mardia (1980), rejecting the assumption of multivariate normality in the study sample. Likewise, the results of the Kolmogorov–Smirnov test with Lilliefors correction yielded values that reject the normal distribution of the *general index* (D = 0.12, $p < 0.001$). The global reliability coefficients of the EEDocente ($\alpha = 0.92$, $\rho = 0.92$ and $\omega = 0.93$) meet the quality criteria established *a priori*. Likewise, the three subscales meet the quality criteria $\alpha \geq 0.70$, $\rho \geq 0.70$. However, regarding the McDonald *Omega* coefficient (ω), subscales 1 and 2 [*Course Organization* (F1) and *Quality of Teaching* (F2)] present values below the $\omega \geq 0.80$ criterion. Table 4 shows the values obtained from the general reliability coefficient and by subscale of the EEDocente.

3.2. Confirmatory factor analysis

The fit indices estimated using the WLS ($\chi^2 = 251.21$; $df = 87$, $p = 0.000$; CFI = 0.868; TLI = 0.841; GFI = 0.936; NNFI = 0.814; RMSEA = 0.034; SRMR = 0.057) and WLSMV ($\chi^2 = 52.80$, $df = 87$, $p = 0.999$, CFI = 1.0, TLI = 1.0, GFI = 0.999, RMSEA = 0.000, SRMR = 0.21) were adequate for the three-factor model of the EEDDocente. In turn, the factors presented on average moderate correlations among themselves (0.64) with correlation coefficients ranging between 0.58 and 0.76. In addition, the standardized factor loadings of the three-factor model showed significant and adequate values (see Figure 1).

3.3. Factorial invariance

Table 5 shows the results of the adjustment of the factorial invariance parameters of the three-factor model of the EEDDocente based on the *School stage* and by *Knowledge Area*. It is shown that the three-factor model of teacher performance from the perception of the students was adequate for the groups according to the *School stage* (*Basic Stage, Disciplinary, Stage and Terminal Stage*) and by *Knowledge Area* (*Accounting and Administrative Sciences, Legal Sciences*, and

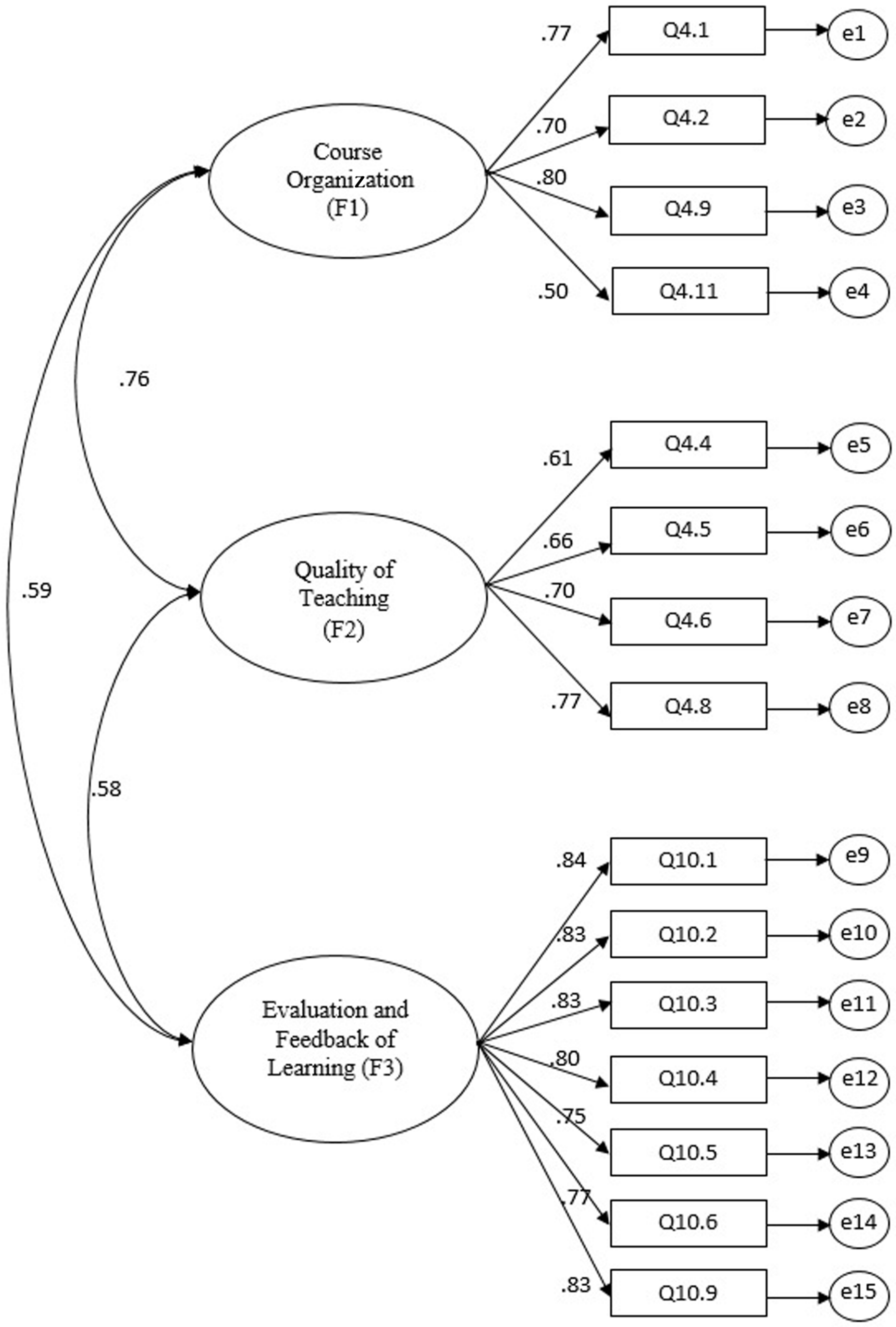

FIGURE 1
Three-factor first-order CFA model of EEDDocente.

Social Sciences). The *Configural* invariance model presented a good fit for all study groups. In particular, the differences between the *Weak*, *Strong* and *Strict* models, both for the groups based on *School stage* and *Knowledge Area*, meet the cut-off criteria (ΔCFI <0.010, ΔRMSEA ≤0.015) (Chen, 2007; Dimitrov, 2010; Putnick and Bornstein, 2016). With the differences obtained between the *Weak* (ΔRMSEA = −0.002 and ΔCFI = −0.001), *Strong* (ΔRMSEA = −0.001 and ΔCFI = −0.002) and Strict (ΔRMSEA = 0.000 and ΔCFI = −0.005) models for the groups depending on the *School stage* and the *Weak* (ΔRMSEA = −0.002) models and ΔCFI = −0.001, *Strong* (ΔRMSEA = −0.002 and ΔCFI = −0.002) and *Stric* (ΔRMSEA = 0.002 and ΔCFI = −0.008) for the groups depending on the educational programs by *Knowledge Area*, factorial invariance is verified.

4. Discussion

The development and validation of the EEDocente represents an important contribution to the study and measurement of teacher performance from the perspective of students (Shevlin et al., 2000;

TABLE 5 Fit indices to evaluate the factorial invariance by *school stage* and *area of knowledge* of the three-factor model of the EEDDocente.

Model	χ^2	*df*	CFI	ΔCFI	TLI	ΔTLI	RMSEA	ΔRMSEA	SRMR	BIC	ΔBIC
School stage											
Configural	812.09	261	0.954	---	0.944	---	0.062	---	0.033	49935.1	---
Week	852.58	285	0.953	−0.001	0.948	0.003	0.060	−0.002	0.043	49806.5	−137.15
Strong	902.89	309	0.950	−0.002	0.949	0.002	0.059	−0.001	0.044	49677.9	−127.34
Strict	989.25	339	0.946	−0.005	0.949	0.000	0.059	0.000	0.047	49543.7	−135.68
Area of knowledge											
Configural	816.14	261	0.950	---	0.940	---	0.065	---	0.034	50125.3	---
Week	856.90	285	0.949	−0.002	0.943	0.003	0.063	−0.002	0.043	49981.6	−135.09
Strong	907.61	309	0.946	−0.002	0.945	0.002	0.062	−0.001	0.045	49858.7	−125.13
Strict	1029.83	333	0.938	−0.008	0.942	−0.003	0.063	0.002	0.052	49779.2	−97.59

Whittington, 2001; Campbell et al., 2005; Richardson, 2005). In particular, this study provides evidence of reliability, internal structure, and factorial invariance that allow for further comparative studies and thus evidence-based decision-making. Contrary to high-stakes assessments, the use of this type of assessment for the purpose of performance improvement and continuous teacher training is a rare practice, but vital for the improvement of classroom education in all education systems around the world. By way of discussion, the most relevant findings of the study are presented and contrasted with the results of other similar studies.

In particular, the reduced version ($K=15$) of the EEDDocente complies with the psychometric quality criteria of reliability and internal structure. The global reliability coefficients of the EEDDocente meet the cut-off criteria ($\alpha=0.92$, $\rho=0.92$ and $\omega=0.93$), and the reliability coefficients per subscale are very close to what was expected. Likewise, with the results of the CFA obtained, the three-factor structure proposed by the Coordination of teacher evaluation of the FCAyS is corroborated (Henríquez et al., 2017, 2018; Henríquez and Arámburo, 2021). The multidimensional model of three factors with 15 items presents adequate factor loadings (between 0.50 and 0.84) and an acceptable. With this, the structure of the EEDDocente, which addresses some of the most relevant teaching competencies throughout the educational levels, endorses and consolidates its underlying theoretical model. This is consistent with other studies of similar instruments that present similar theoretical dimensions in their structure (Marsh, 1984, 1993, 2007; Feldman, 1988; Centra, 1993; Feldman, 1993; Braskamp and Ory, 1994; Fink, 2008; Silva et al., 2014; Irigoyen et al., 2016; Bazán-Ramírez and Velarde-Corrales, 2021; Bazán-Ramírez et al., 2021). It is important to mention that the items eliminated do not affect the interpretation of the construct, maintaining the three basic dimensions of the EEDDocente defined at the beginning by the design committee. Likewise, with a smaller scale, the time and possible problems related to the average number of times a regular student of the FCAyS must answer the EEDDocente per semester are reduced.

Added to this, the study provides new findings on factorial invariance depending on the School stage and *Educational programs* in the *Knowledge Areas* of *Accounting and Administrative Sciences*, *Legal Sciences*, and *Social Sciences* in samples of university students. The *Configural* invariance model presented a good fit for all study groups, and the differences in the parameters of the *Weak* and *Strong* models are adequate. This guarantees that the EEDDocente can be considered on the same scale for the different groups under study and confirms that the three-factor model measures in the same way in all of them (Vandenberg and Lance, 2000; Wang and Wang, 2012). In addition, the Bayesian Information Criterion (BIC) presented a sequential reduction, which can be interpreted as a sign of equivalence between the samples (Cheung and Rensvold, 2002). In this regard, Chen (2007) mentions that the RMSEA and SRMR tend to reject invariant models when the sample size is not equal between the groups, so it is advisable to use the CFI as the main criterion to establish invariance.

It must be recognized that the main limitation of the study has to do with the fact that, although the student samples are large, they are not equitable between the study groups. In particular, it is important to remember that there is a difference greater than 100 individuals between the groups of the *School stage* of the *Basic stage* ($N=632$) and the *Disciplinary stage* ($N=816$), and that this difference increases when compared with the number of students participating in the *Terminal stage* ($N=392$). The same happens with the number of participants in the educational programs by *Area of Knowledge*, where 366 students from the *Knowledge Area* of *Legal Sciences* participated, and in *Accounting and Administrative Sciences* and *Social Sciences* almost twice as many participated ($N=807$ and $N=676$ respectively).

5. Conclusion

By way of conclusion, it can be said that the findings derived from the reliability analysis and the CFA provide evidence that supports an adequate adjustment of the three-factor structural model [*Class Organization* (F1), *Teaching quality* (F2), and *Assessment and Feedback on learning* (F3)] of the reduced version ($K=15$) of the EEDDocente to evaluate teaching performance throughout the *School stage* (*Basic stage*, *Disciplinary stage*, and *Terminal stage*) and the *Areas of knowledge* (*Accounting and Administrative Sciences, Legal Sciences*, and *Social Sciences*). In addition, factorial invariance analysis based on the *School stage* and the *Educational programs* by *Areas of Knowledge* in samples of university students show an adequate adjustment of the *Configural* model, and the differences in the parameters of the *Weak*, *Strong*, and *Strict* models. These results indicate that none of the study groups presents a systematic tendency to answer the items higher or lower than the rest of the groups (Vandenberg and Lance, 2000; Meredith and Teresi, 2006; Wang and Wang, 2012), providing evidence to carry out comparative studies. With all this, it is guaranteed that the EEDDocente complies with the standards of reliability, internal structure validity and invariance, and its use as a brief and easy-to-administer instrument is supported, presenting an important contribution to the study and measurement of teaching from the students' perspective. It is recommended for future research ensure the equivalence of the samples of the study groups to favor the analysis of the metric invariance and factorial invariance of the EEDDocente and to carry out comparative and predictive studies. It is also important to consider the application of the EEDDocente in other schools and universities in order to have a tool for brief application with the purpose of providing relevant information at the end of each school stage, based on the opinion of the students, on teaching performance, favoring continuous training and decision making to improve the effectiveness and quality of teaching.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The protocol and procedure of the evaluation strategy on which the present study was based were approved by the FCAySUABC administration and supervised by the FCAyS-UABC Teaching Evaluation Coordination, in accordance with the institutional norms in force regarding research ethics. For the application of the EEDDocente, a group of students was trained as applicators of the instrument and the voluntary participation of the students of each educational program of the FCAyS-UABC was requested, who were previously informed about the objectives and procedures of the study and about the confidentiality, safeguarding and use of the information obtained based on their answers.

Author contributions

PH and JP-M contributed to the idea of research, its conceptualization, implementation, and methodology, were in charge of writing the manuscript, and also contributed to the analysis and interpretation of data, revising the English version, and writing in Frontiers format. CC provided support in the search for additional bibliographic information, was responsible for drafting the introduction to the manuscript, and reviewed the writing style of the article. JZ contributed to the design and implementation of the EEDDocente, and also provided elements for the delimitation of the participants. All authors contributed to the article and approved the submitted version.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Alcaraz-Salarirche, N. (2015). *Aproximación Histórica a la Evaluación Educativa: De la Generación de la Medición a la Generación Ecléctica [Historical Approach to Educational Evaluation: From the Measurement Generation to the Eclectic Generation]. Revista Iberoamericana de Evaluación Educativa. 8:1*. Available at: https://dialnet.unirioja.es/servlet/articulo?codigo=5134142 (Accessed January 21, 2023).

Aleamoni, L. (1999). Student rating myths versus research facts from 1924 to 1998. *J. Pers. Eval. Educ.* 13, 153–166. doi: 10.1023/a:1008168421283

American Educational Research Association (2018). *American Psychological Association y National Council on Measurement Education [AERA, APA, NCME]*. Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Aravena-Gaete, M. E., and Gairín, J. (2021). Evaluación del desempeño docente: Una mirada desde las agencias certificadoras. [Evaluation of teacher performance: a look from the certification agencies]. *Prof. Rev. Currículum Form. Prof.* 25, 297–317. doi: 10.30827/profesorado.v25i1.8302

Arreola, R. A. (2007). *Developing a comprehensive faculty evaluation system: A guide to designing, building, and operating large-scale faculty evaluation systems (3rd)*. Bolton, MA: Anker.

Bazán-Ramírez, A., and Velarde-Corrales, N. (2021). Auto-reporte del estudiantado en criterios de desempeño didáctico en clases de Psicología [students self-report within didactic performances criteria in psychology classes]. *J. Behav. Health Soc.* 13, 22–35. doi: 10.22201/fesi.20070780e.2021.13.1.78071

Bazán-Ramírez, A., Pérez-Morán, J. C., and Bernal-Baldenebro, B. (2021). Criteria for teaching performance in psychology: invariance according to age, sex, and academic Stage of Peruvian students. *Front. Psychol.* 12:764081. doi: 10.3389/fpsyg.2021.764081

Benton, S. L. (2018). Best practices in the evaluation of teaching. *Best Pract. Eval. Teach.* 69, 1–18.

Bleiberg, J. (2023). "Revisiting teacher evaluation a decade after reforms" in *International encyclopedia of education*. eds. R. Tierney, F. Rizvi and K. Ercikan. *4th* ed (Amsterdam, Netherlands: Elsevier)

Boring, A., Ottoboni, K., and Stark, P. B. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *Sci. Open Res.* 1, 1–11. doi: 10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1

Boštjančič, E., Komidar, L., and Johnson, R. B. (2018). Factorial validity and measurement invariance of the slovene version of the cultural intelligence scale. *Front. Psychol.* 9:1499. doi: 10.3389/fpsyg.2018.01499/full

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

Brown, J. D. (2012). "Classical test theory" in *The Routledge handbook of language testing measurement*. eds. G. Fulcher and F. Davidson (Abingdon: Routledge)

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. New York: Guilford Publications.

Browne, M. W., and Cudeck, R. (1993). "Alternative ways of assessing model fit, testing structural equation models" in *Testing estructural models*. eds. K. A. Bollen and J. S. Long (Thousand Oaks, California: Sage)

Byrne, B. M., Shavelson, R. J., and Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: the issue of partial measurement invariance. *Psychol. Bull.* 105, 456–466. doi: 10.1037/0033-2909.105.3.456

Calatayud, M. (2021). Evaluación docente y mejora profesional. Descubrir el Encanto de su complicidad. [teacher evaluation and professional improvement. Discover the charm of their complicity]. *Rev. Iberoam. Eval. Educ.* 14, 87–100. doi: 10.15366/riee2021.14.1.005

Camacho, I. (2022). El desempeño docente y su implicación en la enseñanza. Teacher performance and their involvement in teaching. *Form. Estrat.* 6, 105–120.

Campbell, H. E., Steiner, S., and Gerdes, K. (2005). Student evaluations of teaching: how you teach and who you are. *J. Public Aff. Educ.* 11, 211–231. doi: 10.1080/15236803.2005.12001395

Carpio, C., Pacheco, V., Canales, C., and Flores, C. (1998). Comportamiento inteligente y juegos de lenguaje en la enseñanza de la psicología [Intelligent behavior and language games in the teaching of psychology]. *Acta Comport* 6, 47–60.

Cashin, W. E., and Downey, R. G. (1992). Using global student rating items for summative evaluation. *J. Educ. Psychol.* 84, 563–572. doi: 10.1037/0022-0663.84.4.563

Centra, J. A. (1993). *Reflective faculty evaluation: Enhancing teaching and determining faculty effectiveness*. San Francisco, CA: Jossey-Bass.

Chan, W. M. (2018). Teaching in HIGHER education: students' perceptions of effective teaching and good teachers. *Soc. Sci. Educ. Res. Rev* 5, 40–58.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Struct. Equ. Model.* 14, 464–504. doi: 10.1080/10705510701301834

Chen, Y., and Hoshower, L. B. (2003). Student evaluation of teaching effectiveness: an assessment of student perception and motivation. *Assess. Eval. High. Educ.* 28, 71–88. doi: 10.1080/02602930301683

Cheung, G. W., and Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Struct. Equ. Model.* 9, 233–255. doi: 10.1097/NNR.0b013e3182544750

Ching, G. (2018). A literature review on the student evaluation of teaching: an examination of the search, experience, and credence qualities of SET. *High. Educ. Eval. Dev.* 12, 63–84. doi: 10.1108/HEED-04-2018-0009

Cisneros-Cohernour, E., and Stake, R. (2010). La evaluación de la docencia en educación superior: de evaluaciones basadas en opiniones de estudiantes a modelos por competencias [the evaluation of teaching superior education: from evaluations based on opinions of students in models of competences]. *Rev. Iberoam. Eval. Educ.* 3, 218–231.

Cohen, P. A. (1981). Student ratings of instruction and student achievement: a meta-analysis of multisection validity studies. *Rev. Educ. Res.* 51, 281–309. doi: 10.2307/1170209

Cohen, P. A. (1983). Comment on a selective review of the validity of student ratings of teaching. *J. High. Educ.* 54, 448–458. doi: 10.2307/1981907

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika* 16, 297–334. doi: 10.1007/BF02310555

Cronbach, L. J. (1988). Internal consistency of tests: analyses old and new. *Psychometrika* 53, 63–70. doi: 10.1007/BF02294194

Cruz Ávila, M. (2007). *Una propuesta para la evaluación del profesorado universitario*. A proposal for the evaluation of university teaching staff.Trabajo de investigación para la obtención del grado de doctor. Universitat Autonoma de Barcelona. Available at: https://www.tdx.cat/bitstream/handle/10803/5285/mca1de1.pdf (Accessed January 16, 2023).

Dallal, G., and Wilkinson, L. (1986). An analytic approximation to the distribution of Lilliefors's test statistic for normality. *Am. Stat.* 40, 294–296. doi: 10.1080/00031305.1986.10475419

Dimitrov, D. M. (2010). Testing for factorial invariance in the context of construct validation. *Meas. Evaluat. Counsel. Dev.* 43, 121–149. doi: 10.1177/0748175610373459

Estrada, L. A., Yglesias, L. A., Miranda, A. E., Díaz, J. K., and Díaz, S. M. (2019). Propiedades Psicométricas de un Cuestionario sobre Evaluación del Desempeño Docente Universitario desde la Percepción del Estudiante [psychometric properties of a questionnaire on evaluation of university teaching performance from the perception of the student]. *Rev. Investig. Estadíst.* 2, 92–102.

Feldman, K. A. (1976). The superior college teacher from the students' view. *Res. High. Educ.* 5, 243–288. doi: 10.1007/BF00991967

Feldman, K. A. (1988). Effective college teaching from the students' and faculty´s view: matched or mismatched priorities? *Res. High. Educ.* 28, 291–329. doi: 10.1007/BF01006402

Feldman, K. A. (1989a). Instruccional effectiveness of college teachers as judged by teacher themselves, current and former students, colleagues, administrators, and external (neutral) observers. *Res. High. Educ.* 30, 137–194.

Feldman, K. A. (1989b). The association between student ratings of specific instructional dimensions and student achievement: refining and extending the synthesis of data from multisectional validity studies. *Res. High. Educ.* 30, 583–645.

Feldman, K. A. (1990). An afterword for "the association between student ratings of specific instructional dimensions and student achievement": refining and extending the synthesis of data from multisection validity students. *Res. High. Educ.* 31, 315–317.

Feldman, K. A. (1992). College students' views of male and female college teachers: part I- evidence from the social laboratory experiments. *Res. High. Educ.* 33, 317–375.

Feldman, K. A. (1993). College students' views of male and female teachers: part II-evidence from students' evaluations of their classroom teachers. *Res. High. Educ.* 34, 151–211.

Fink, D. (2008). "Evaluating teaching: a new approach to an old problem" in *To improve the academy: Resources for faculty, instructional, and organizational development*. eds. S. Chadwick-Blossey and D. R. Robertson (Hoboken, NJ: Jossey-Bass)

García, J. M. (2000). "Las dimensiones de la efectividad docente, validez y confiabilidad de los cuestionarios de evaluación de la docencia: síntesis de investigación internacional [the dimensions of teaching effectiveness, validity and reliability of teacher evaluation questionnaires: international research synthesis]" in *Evaluación de la docencia*. eds. M. Rueda and Y. F. Díaz-Barriga (Mexico: Paidós), 41–62.

García, J. M. (2014). *¿Para qué sirve la evaluación de la docencia? Un estudio exploratorio de las creencias de los estudiantes [how useful is the evaluation of teaching? An exploratory study on students' beliefs]*. Education policy analysis archives, 22. pp. 1–20. Available at: https://www.redalyc.org/pdf/2750/275031898010.pdf (Accessed January 14, 2023).

García-Gómez-Heras, S., Gil-Madrona, P., Hernandez-Barrera, V., Lopez-de-Andres, A., and Carrasco-Garrido, P. (2017). Importance of university teacher behaviour in the faculty of health science. *Aust. Med. J.* 10, 800–810. doi: 10.21767/AMJ.2017.3128

García-Olalla, A., Villa-Sánchez, A., Aláez, M., and Romero-Yesa, S. (2022). Aplicación y resultados de un sistema para evaluar la calidad de la docencia universitaria en una década de experimentación [Implementation and results of a system to evaluate the quality of University teaching in a decade of experimentation]. *Rev. Investig. Educ.* 40, 51–68. doi: 10.6018/rie.401221

Gazeloglu, C., and Greenacre, Z. A. (2020). Comparison of weighted least squares and robust estimation in structural equation modeling of ordinal categorical data with larger sample sizes. *Cumhuriyet Sci. J.* 41, 193–211. doi: 10.17776/csj.648054

Gómez, L. F., and Valdés, M. G. (2019). The evaluation of teacher performance in higher education. *Propósitos Represent.* 7, 479–515. http://dx.doi.org/10.20511/pyr2019.v7n2.255.

Gu, R., Wang, H. N., and Lou, L. S. (2021). Optimization and application of data analysis strategy for college students' evaluation of teaching. *J. Zhejiang Univ. Tech.* 20, 201–207.

Hair, J. F., Black, W. C., Babin, B. J., and Anderson, R. E. (2019). *Multivariate Data Analysis*. Hampshire: Cengage Learning EMEA.

Henríquez, P., and Arámburo, V. (2021). Evaluación del desempeño docente por áreas de conocimiento: El Caso de la Facultad de Ciencias Administrativas y Sociales de la Universidad Autónoma de Baja California, México. [evaluation of teaching performance by areas of knowledge: the case of the Faculty of Administrative and Social Sciences of the Autonomous University of Baja California, Mexico]. *Act. Investig. Educ.* 21, 1–20. doi: 10.15517/aie.v21i3.46294

Henríquez, P., Arámburo, V., and Boroel, B. (2018). *Análisis de las estrategias de enseñanza según áreas de conocimiento en el nivel educativo superior: Percepciones de estudiantes de la Facultad de Ciencias Administrativas y Sociales (FCAYS), Universidad Autónoma de Baja California (UABC), México. [Analysis of teaching strategies according to areas of knowledge at the higher education level: Perceptions of students from the School of Administrative and Social Sciences (FCAYS), Autonomous University of Baja California (UABC), Mexico]*. Compendio Investigativo de Academic Journals Celaya 2018, 14. pp. 2320–2325.

Henríquez, P., Arámburo, V., and Dávila, E. (2017). *Percepción de los estudiantes universitarios acerca de las estrategias pedagógicas y de evaluación del aprendizaje utilizadas por sus profesores: el Caso de la FCAYS de la UABC. [perception of university students about the pedagogical and learning assessment strategies used by their professors: the case of the FCAYS at UABC]*. In Memorias electrónicas del XIV Congreso Nacional de Investigación Educativa, COMIE. San Luis Potosí: México. 1–14. Available at: https://www.comie.org.mx/congreso/memoriaelectronica/v14/doc/1956.pdf (Accessed January 24, 2023).

Henríquez, P., Pérez-Morán, J. C., Cid, C. D., and Zamora, J. E. (2023). *Reporte técnico de las propiedades psicométricas de estructura factorial e invarianza de la Escala de Evaluación del Desempeño Docente (EEDDocente) de la FCAyS*. Technical report on the psychometric properties of factorial structure and invariance of the Teaching Performance Evaluation Scale (EEDDocente) of the FCAyS. Documento interno [internal document].

Hildebrand, M. (1973). The character and skills of the effective professor. *The Journal of Higher Education.* 44, 41–50. doi: 10.2307/1980624

Hildebrand, M., Wilson, R.C., and Dienst, E.R. (1971). *Evaluating university teaching*. Center for Research and Development in Higher Education. Berkeley, CA.

Hirschfeld, G., and Von-Brachel, R. (2014). Multiple-group confirmatory factor analysis in R–a tutorial in measurement invariance with continuous and ordinal indicators. *Pract. Assess. Res. Eval.* 19, 1–12. doi: 10.7275/qazy-2946

Hornstein, H. A. (2017). Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance. *Cogent Educ.* 4, 1–8. doi: 10.1080/2331186x.2017.1304016

House, E. R. (1998). Acuerdos institucionales para la evaluación. [Institutional arrangements for evaluation]. *Perspectivas* 28, 123–131.

Hu, L. T., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct. Equ. Model A Multidiscip. J.* 6, 1–55. doi: 10.1080/10705519909540118

International Test Commission. (2013). *ITC Guidelines on Test Use. ITC-G-TU-20131008*. Available at: https://www.intestcom.org/files/guideline_test_use.pdf. (Accessed January 15, 2023).

Irigoyen, J. J., Jiménez, M., and Acuña, K. (2016). Discurso didáctico e interacciones sustitutivas en la enseñanza de las ciencias [Didactic discourse and substitute interactions in teaching Sciences]. *Enseñ. Investig. Psicol* 21, 68–77.

Jöreskog, K. G., and Sörbom, D. (1979). *Advances in factor analysis and structural equation models*. Cambridge, MA: ABT Books.

Jorgensen, T. D., Mansolf, M., and Enders, C. K. (2022). *Pooled Score Tests for Multiply Imputed Data Using semTools*. Available at: https://rdrr.io/cran/semTools/man/lavTestScore.mi.html (Accessed January 23, 2023).

Kane, M. (2006). "Validation" in *Educational measurement*. ed. R. L. Brennan (New Jersey: National Council on Measurement in Education and American Council on Education)

Katz, M. H. (2006). *Multivariable analysis* (*2nd Edn.*). Cambridge: Cambridge University Press.

Kline, R. B. (2015). *Principles and practice of structural equation modeling. 4th Edn*. New York: Guilford publications.

König, J., Ligtvoet, R., Klemenz, S., and Rothland, M. (2017). Effects of opportunities to learn in teacher preparation on future teachers' general pedagogical knowledge: analyzing program characteristics and outcomes. *Stud. Educ. Eval.* 53, 122–133. doi: 10.1016/j.stueduc.2017.03.001

Lera, M., León, J., and Ruiz, P. (2021). Adaptation of the teacher efficacy scale to measure effective teachers' educational practices through students' ratings: a multilevel approach. *Psicothema* 33, 509–517. doi: 10.7334/psicothema2020.262

Liebowitz, D. (2021). Teacher evaluation for accountability and growth: should policy treat them as complements or substitutes? *Labour Econ.* 71:102024. doi: 10.1016/j.labeco.2021.102024

Liebowitz, D. (2022). Teacher evaluation for growth and accountability: under what conditions does it improve student outcomes? *Harv. Educ. Rev.* 92, 533–565. doi: 10.17763/1943-5045-92.4.533

Luna, E., and Torquemada, A. (2008). Los cuestionarios de evaluación de la docencia por los alumnos: balance y perspectivas de su agenda. [questionnaires for the evaluation of teaching by students: balance and perspectives of their agenda]. *Rev. Electrón. Investig. Educ. Espec.* 10, 1–11.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika* 57, 519–530. doi: 10.1093/biomet/57.3.519

Mardia, K. V. (1980). *"Tests for univariate and multivariate normality" in handbook of statistics*. Amsterdam, Netherlands: Elsevier.

Márquez, L., and Madueño, M. L. (2016). Propiedades psicométricas de un instrumento Para apoyar el proceso de evaluación del docente universitario. [psychometric properties of an instrument to support the evaluation process of the university professor]. *Rev. Electrón. Investig. Educ.* 18, 53–61.

Marsh, H. W. (1984). Students' evaluations of university teaching: dimensionality, reliability, validity, potential biases, and utility. *J. Educ. Psychol.* 76, 707–754. doi: 10.1037/0022-0663.76.5.707

Marsh, H. W. (1986). Applicability paradigm: students' evaluations of teaching effectiveness in different countries. *J. Educ. Psychol.* 78, 465–473. doi: 10.1037/0022-0663.78.6.465

Marsh, H. W. (1987). Students' evaluations of university teaching: research findings, methodological issues, and directions for future research. *Int. J. Educ. Res.* 11, 253–388. doi: 10.1016/0883-0355(87)90001-2

Marsh, H. W. (1993). Multidimensional students' evaluations of teaching effectiveness. *J. High. Educ.* 64, 1–18. doi: 10.1080/00221546.1993.11778406

Marsh, H. W. (2007). "Students' evaluations of university teaching: dimensionality, reliability, validity, potential biases and usefulness" in *The scholarship of teaching and learning in higher education: An evidence-based perspective*. eds. R. P. Perry and J. C. Smart (Berlin, Germany: Springer Science and Business Media)

Marsh, H. W., and Dunkin, M. J. (1997). "Students' evaluations of university teaching: a multidimensional perspective" in *Effective teaching in higher education: Research and practice*. eds. R. P. Perry and J. C. Smart (New York: Agathon), 241–320.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychol. Methods* 23, 412–433. doi: 10.1037/met0000144

Meredith, W., and Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Med. Care* 44, S69–S77. doi: 10.1097/01.mlr.0000245438.73837.89

Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am. Psychol.* 50, 741–749. doi: 10.1037/0003-066X.50.9.741

Milfont, T. L., and Fischer, R. (2010). Testing measurement invariance across groups: applications in cross-cultural research. *Int. J. Psychol. Res.* 3, 111–130. doi: 10.21500/20112084.857

Mitchell, K., and Martin, J. (2018). Gender Bias in Student Evaluations. PS. *Polit. Sci. Polit.* 51, 648–652. doi: 10.1017/S104909651800001X

Mohammadi, M. (2021). Dimensions of teacher performance evaluation by students in higher education. *Shanlax Int. J. Educ.* 9, 18–25. doi: 10.34293/education.v9i2.3673

Nájera-Catalán, H. E. (2019). Reliability, population classification and weighting in multidimensional poverty measurement: a Monte Carlo study. *Soc. Indic. Res.* 142, 887–910. doi: 10.1007/s11205-018-1950-z

Nasser-Abu, F. (2017). Teaching in higher education: good teaching through students' lens. *Stud. Educ. Eval.* 54, 4–12. doi: 10.1016/j.stueduc.2016.10.006

Navarro, C., and Ramírez, M. (2018). Mapeo sistemático de la literatura sobre evaluación docente (2013-2017). [systematic mapping of the literature on teacher evaluation (2013-2017)]. *Educ. Pesqui.* 44, 1–23. doi: 10.1590/S1678-4634201844185677

Putnick, D. L., and Bornstein, M. H. (2016). Measurement invariance conventions and reporting: the state of the art and future directions for psychological research. *Dev. Rev.* 41, 71–90. doi: 10.1016/j.dr.2016.06.004

Raykov, T., and Marcoulides, G. A. (2019). Thanks coefficient alpha, we still need you! *Educ. Psychol. Meas.* 79, 200–210. doi: 10.1177/0013164417725127

Revelle, W., and Revelle, M. W. (2015). Package 'psych'. *Comprehens. R Arch. Netw.* 337:338.

Reyes, E., Luna, E., and Caso, J. (2020). Evidencias de validez del Cuestionario de Evaluación de la Competencia Docente Universitaria. [evidence of validity of the university teaching competence assessment questionnaire]. *Perfiles Educ.* 42, 106–122. doi: 10.22201/iisue.24486167e.2020.169.58931

Richardson, J. T. (2005). Instruments for obtaining student feedback: a review of the literature. *Assess. Eval. High. Educ.* 30, 387–415. doi: 10.1080/02602930500099193

Romero, T., and Martínez, A. (2017). Construcción de instrumentos de evaluación del desempeño docente universitario desde una perspectiva cualitativa. [Construction of evaluation instruments of university teaching performance from a qualitative perspective]. *Rev. Univ. Caribe* 18, 34–43. doi: 10.5377/ruc.v18i1.4800

Rosseel, Y. (2012). Lavaan: an R package for structural equation modeling. *J. Stat. Softw.* 48, 1–36.

Salazar, J. (2008). Diagnóstico Preliminar sobre Evaluación de la Docencia Universitaria. Una Aproximación a la Realidad en las Universidades Públicas y/o Estatales de Chile. [Preliminary Diagnosis on Evaluation of University Teaching. An Approach to Reality in Public and/or State Universities in Chile]. *Rev. Iberoam. Eval. Educ.* 1:e3.

Schellhase, K. C. (2010). The relationship between student evaluation of instruction scores and faculty formal educational coursework. *Athl. Train. Educ. J.* 5, 156–164. doi: 10.4085/1947-380X-5.4.156

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., and King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: a review. *J. Educ. Res.* 99, 323–338. doi: 10.3200/JOER.99.6.323-338

Seivane, M. S., and Brenlla, M. E. (2021). Evaluación de la calidad docente universitaria desde la perspectiva de los estudiantes. [evaluation of university teaching quality from the students' perspective]. *Rev. Iberoam. Evalu. Educ.* 14, 35–46. doi: 10.15366/riee2021.14.1.002

Shevlin, M., Banyard, P., Davies, M., and Griffiths, M. (2000). The validity of student evaluation of teaching in higher education: love me, love my lectures? *Assess. Eval. High. Educ.* 25, 397–405. doi: 10.1080/713611436

Silva, L. A. (2023). Congruencia entre la práctica docente y la evaluación de la docencia por parte del estudiantado en educación superior: estudio de casos en la Universidad Veracruzana. [congruence between teaching practice and the evaluation of teaching by students in higher education: case study at the Universidad Veracruzana]. *Rev. Educ.* 47, 114–128. doi: 10.15517/revedu.v47i1.51978

Silva, H., Morales, G., Pacheco, V., Camacho, A., Garduño, H., and Carpio, C. (2014). Didáctica como conducta: una propuesta para la descripción de las habilidades de enseñanza [Didactic as behavior: a proposal for the description of teaching skills]. *Rev. Mexic. Anál. Conduc.* 40, 32–46. doi: 10.5514/rmac.v40.i3.63679

Silva, J., Solís, P., Huaman, J., and Quispe, F. (2022). La evaluación formativa en el desempeño docente universitario: Revisión sistemática de literatura. [Formative evaluation in university teaching performance: Systematic literature review]. *Tecnohum. Rev. Cien.* 2, 1–14. doi: 10.53673/th.v2i4.177

Spooren, P., Brockx, B., and Mortelman, D. (2013). On the validity of student evaluation of teaching: the state of the art. *Rev. Educ. Res.* 83, 598–642. doi: 10.3102/0034654313496870

Stroebe, W. (2016). *Student evaluations of teaching: no measure for the TEF*. Times Higher Education. Available at: https://www.timeshighereducation.com/comment/student-evaluations-teaching-no-measure-tef (Accessed July 3, 2023).

Stroebe, W. (2020). Student evaluations of teaching encourages poor teaching and contributes to grade inflation: a theoretical and empirical analysis. *Basic Appl. Soc. Psychol.* 42, 276–294. doi: 10.1080/01973533.2020.1756817

R Core Team. (2022). *Writing R extensions. R Foundation for Statistical Computing*, pp. 1–208. Available at: https://cran.r-project.org/doc/manuals/R-exts.html (Accessed July 4, 2023).

Theall, M., Abrami, C., and Lisa, A. (2001). *The student ratings debate: Are they valid? How can we best use them*. San Francisco, California: Jossey Bass Press.

Tomás-Folchy, M., and Durán-Bellonch, M. (2017). Comprendiendo los factores que afectan la transferencia de la formación permanente del profesorado. Propuestas de mejora [understanding the factors affecting transfer of university teachers' permanent training. Proposals for improvement]. *Rev. Electrón. Interuniv. Form. Prof.* 20, 145–157. doi: 10.6018/reifop/20.1.240591

Torquemada, A. (2022). *Evaluación, desarrollo, innovación y futuro de la docencia universitaria. [evaluation, development, innovation and future of university teaching]*. De la Red Iberoamericana de Investigadores en Evaluación de la Docencia. Perfiles educativos, 44. pp. 200–204.

Tuncel, S. D. (2009). Determining effective teacher behavior contributing to students' academic success. *Int. J. Phys. Educ.* 1, 15–18.

Vaillant, D. (2016). El fortalecimiento del desarrollo profesional docente: una mirada desde Latinoamérica. [the strengthening of teacher professional development: a look from Latin America]. *J. Supranational Policies Educ.* 5:5-21. doi: 10.15366/jospoe2016.5

Vance, R. I., Winne, P. S., and Wright, E. S. (1983). A longitudinal examination of rater and ratee effects in performance ratings. *Pers. Psychol.* 36, 609–620. doi: 10.1111/j.1744-6570.1983.tb02238.x

Vandenberg, R. J., and Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organ. Res. Methods* 3, 4–70. doi: 10.1177/109442810031002

Wang, D. F., and Guan, L. (2017). Higher education quality evaluation from the perspective of students: theoretical construction and reflection. *J. Nat. Inst. Educ. Admin.* 5:75. doi: 10.3969/j.issn.1672-4038.2017.05.005

Wang, J., and Wang, X. (2012). *Structural equation modeling applications using Mplus*. Chichester: John Wiley & Sons Ltd..

Whittington, L. (2001). Detecting good teaching. *J. Public Aff. Educ.* 7, 5–8. doi: 10.1080/15236803.2001.12023490

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., et al. (2019). Welcome to the Tidyverse. *J. Open Source Softw.* 4:1686. doi: 10.21105/joss.01686

Williams, K., and Hebert, D. (2020). Evaluation systems: a literature review on issues and impact. Louisiana educational research association. *Research Issues in Contemporary Education* 5, 42–50.

Zamora, E. (2021). La evaluación del desempeño docente mediante cuestionarios en la universidad: Su legitimidad según la literatura y los requerimientos Para que sea efectiva [the evaluation of teaching performance through questionnaires at the university: its legitimacy according to the literature and the requirements for it to be effective]. *Rev. Actual. Investig. Educ.* 21, 1–23. doi: 10.15517/aie.v21i3.46221

Zhang, Z., and Yuan, K. H. (2016). Robust coefficients alpha and omega and confidence intervals with outlying observations and missing data: methods and software. *Educ. Psychol. Meas.* 76, 387–411. doi: 10.1177/0013164415594658

Zhou, J. L., and Qin, Y. (2018). The basic types of college students' teaching evaluation behavior deviation and its relationship with students' background characteristics. *Fudan Educ. Forum.* 2018:6. doi: 10.3389/fpsyg.2022.1004487

Zhao, L., Xu, P., Chen, Y., and Yan, S. (2022). A literature review of the research on students' evaluation of teaching in higher education. *Front. Psychol.* 13:1004487. doi: 10.3389/fpsyg.2022.1004487