Psychometric perspectives in Educational and Learning Capitals: Development and validation of a scale on student evaluation of teaching in Higher Education

Tarquino Sánchez[a], Alejandro Veas[b], Raquel Gilar-Corbí[b] & Juan-Luis Castejón[b]

[a] National Polytechnic School of the Ecuador. Department of Electronic, Telecommunications and information networks

[b] University of Alicante. Department of Developmental Psychology and Didactics

Abstract:
Student evaluation of teaching is an important topic in the field of education, and different rating scales have been developed in international contexts. However, there have been methodological problems that may lead to ineffective construct measurement. The aim of this study Considering as an extension of the Actiotope Model of Giftedness, the aim of this study was to support an internal structure of a new 32 item instrument in a large public university in Ecuador, using a sample of 6110 students. Data were analysed based on the item response theory, including unidimensional and multidimensional Rasch rating scale models to examine three theory-based constructs. Preference was given to a more precise multidimensional construct with four interrelated domains.

Keywords:
student evaluation of teaching, scale validation, higher education, item response theory

1 Introduction

Student evaluation of teaching (SET) ratings is a generalized practice in almost every institution of higher education (Huybers, 2014; Richardson 2005; Zabaleta, 2007). Regarded as eliciting perceived performance feedback on a range of teacher-and/or course-related aspects, in most universities, SET is used for formative purposes (e.g. feedback for the improvement of instruction) as well as for administrative decision-making (e.g., about recruitment, career progress, and economic incentives) (Linse, 2017).

Considering the educational dynamic context as one of the most relevant factors in any teaching-learning process, the purpose of the present study was to develop a measure of SET from a large public institution in Ecuador. The access to institution-wide student survey data may reveal possible structural or psychological barriers, considering the underlying relations among beliefs, knowledge, and actions involved in the field.

1.1 Theoretical perspectives on SET

The analysis of the teaching process has been raised as an important concern in the last decades which has been related to the diversity of models of teaching. Initially, teacher competency moved from on a series of actions and behaviors (Boice, 1991) to a more complex cognitive activity (Leinhardt & Greeno, 1986) and affections (Berlinger, 1986).

Given these diverse structures, the initial frameworks of effective teaching started with the commonly named theories of expertise, in which the most valuable information was reported through indicators like depth of problem representation, knowledge organization and structure, efficiency of procedures, and metacognitive skills for learning, among others (Glaser, Lesgold, & Lajoie, 1984). However, particularly in higher education, beliefs or views about teaching have also been considered as important constructs in effective teaching (Larsson, 1986). Considering beliefs as "implicit assumptions about students, learning classrooms, and the subject matter" (Pajares, 1992; Pratt, 1997), this framework moves toward a developmental and integrated view of teaching and the learning process as a necessary link to an effective SET.

Within a more dynamic perspective, the theory of transformative learning (Mezirow, 1991) posits an important distinction between individuals' previous beliefs of teaching and what they actually do when they teach. In those situations, the identification of contextual barriers is crucial to determine, as well as the identification of teachers' lack of effective strategies to implement positive assumptions (Schön, 1983). The change towards effective teaching lies in an effective engagement process with support and application of strategies in positive learning conditions.

In this setting, the authors of the present study relied on "the new model of teaching" provided by Saroyan and Amundsen (2001). This ecological framework considers three main elements associated with SET: conceptions or beliefs, knowledge and actions. Moreover, the key element for dynamic formation of the concepts are based on the analyses of the contextual influences. In this latter construct, it includes all external factors which may influence teaching tasks, such as the culture of the university, faculty or department, or instructors' teaching assignments (Saroyan & Amundsen, 2001).

Noting this ecological perspective, in Latin-American higher education institutions, special efforts have been taken to be considered as relevant actors of social development (Arocena & Sutz, 2005). During the twentieth century, important social movements triggered the so- called University Reform Movement (URM) (Ribeiro, 1971), allowing the inclusion of social policies in higher education, in spite of political or military controversies through the past decades. It can be said that this spread of democracy in the higher education system has met the goals of stronger teaching and research standards.

The Council of Ecuadorian Higher Education established the obligatory nature of the evaluation of the teaching in the higher education institutions, both for its organization and promotion, in the Career and Ladder Regulations of the Professor and Researcher of the Higher Education System (CES, 2017). The assessment of the performance of university teachers is an essential component that allows a professor to enter as an Assistant Professor or Associate Professor. The requirements include a score of at least 75% in the performance evaluation during the last two academic periods. Additionally, according to Article 96 of the regulations, members of the academic staff will be dismissed if they have obtained: 1) two consecutive integral evaluation of performance scores of less than 60%; and 2) four integral evaluations of performance scores of less than 60% during their careers. In addition, it establishes that the main titular teachers will be promoted to the next higher level if they comply with other requirements such as having obtained at least 80% on their performance evaluation scores in the last two academic periods (Consejo de Educación Superior [CES], 2017).

1.2 Connecting Educational and Learning Capitals through SET

Recently, synthetic perceptions of achievement levels have been proposed as an effective strategy to overcome analytic strategies (Veas et al., 2018). From the giftedness field, the transformation of talents, gifts, or abilities into achievements is still considered a linear sum of independent variables. This basic assumption is also virtually expressed in graphic representations of models when neatly separated boxes of variables are listed after bullet points. However, usually no information is given about the exact nature of the interplay and the involved processes (Heller et al., 2005).

A synthetic perspective of achievement can be proposed from an extension of the actiotope model of giftedness (Ziegler et al., 2017; Ziegler & Stoeger, 2017) to all achievement levels, as every student constitutes an actiotope with specific resources. Concretely, the influx of exogenous resources from the environment is of particular importance. In this context, Ziegler and Baker (2013) referred environment resources as educational capital, whereas internal resources are considered as learning capital -resources that can be used to promote learning). Within the SET perspective, student's criteria consist on a cultural educational capital, which includes thinking patterns which can facilitate – or hinder- the attainment of learning and educational goals (Ziegler & Baker, 2013, p. 28). For this reason, this resource directly affects to the teachers' view of their own professional development, which can be composed by telic learning capital and actional learning capital (Ziegler et al., 2017; Ziegler et al., 2019). Indeed, evaluation of teaching is effective when educational systems may pro-

mote changes on teachers' goals to activate professional actions, and therefore optimal results.

1.3 Review of SET measures

The instruments normally used to measure students' evaluation of their teachers, programs, and students' satisfaction with their instruction, are the standard rating scales. However, research on SET ratings have not yet provided a clear answer about some questions of their validity (Hornstein, 2017; Marsh, 2007a, b; Spooren, Brockx, & Mortelmans, 2013; Uttl et al., 2017). Interestingly, many of the evaluation instruments have been constructed and validated within the institution itself, and the results of this validation have not always been published nor they have even been tested for their psychometric quality (Richardson, 2005).

Several well-designed and validated SET instruments are available (Spooren et al., 2013). One of these instruments more widely used is the one published by Marsh (1982), and Marsh et al (2009), the Student's Evaluations of Educational Quality (SEEQ). This scale is composed of 35 statements which to each the students respond, using a five-point scale. The scale evaluates nine aspects of teaching: learning value, enthusiasm, organization, group interaction, individual rapport, breadth of coverage, exams and grading, assignments, workload or difficulty. These dimensions have been reproduced in the Confirmatory Factor Analyses, using large samples in different countries, different teacher status, and disciplines (Marsh, 1987; 2007a). Other more recently developed assessment instruments are the Students' Evaluation of Teaching Effectiveness Rating Scale (SETERS) by Toland and De Ayala (2005), the Teaching Proficiency Item Pool (Barnes et al., 2008), the SET37 inventory by Mortelmans and Spooren (2009) and the Teaching Behavior Checklist (Keeley, Furr, & Buskist, 2010).

However, there is no consensus on the number and type of dimensions. This lack of consensus (Apodaca & Grad, 2005; Spooren et al., 2013) is due to conceptual problems related to the lack of a common theoretical framework about what effective teaching is, and methodological problems concerning the measurement of dimensions, as a data-driven process in which different post hoc analytic techniques are used. It seems necessary to use the most common dimensions that are associated with greater teaching effectiveness.

The question concerning construct validity that arises in relation to SEEQ and other instruments is about its unidimensional or multidimensional structure. Researchers agree that there are several dimensions, but it is not clear whether they can be subsumed into a single global dimension. Marsh (1987, 1991a, b, 2007) considers that, although the dimensions of the SEEQ were correlated with each other, they are not represented by a general higher-order factor.

Against the assertion of Abrami, d'Apollonia, and Rosenfield (1997) that the SEEQ dimensions were subsumed by a single construct called "general instructional skill"; based on the evidence from the results of Hierarchical Confirmatory Factor Analysis (HCFA), Marsh (1991b, 2007a) concluded that support for a multidimensional view is strong. Conversely, using HCFA, Cheung (2000) found Evidence of a Single Second-Order Factor in Student Ratings of Teaching. Marsh et al (2009) defended a multidimensional structure of the students' evaluations of university teaching (SETs), on the basis of which measures can be ob-

tained: both of the specific dimensions and of a general factor of the quality of teaching.

The test-retest reliability of students' evaluations is high, even when there is a long period of time between evaluations (Richardson, 2005). There is a high correlation between the scores of students taking different subject-matter taught by the same teacher, but a low relationship between the evaluations given by students taking the same subject-matter taught by different teachers. This suggests that students' evaluations are a function of the teacher's appeal rather than the subject matter (Marsh, 2007). Moreover, evaluations of the same teachers given by successive cohorts of students are highly correlated (Marsh, 2007b; Marsh & Hocevar, 1991; Marsh et al., 2009).

The inter-rater reliability of individuals and the average ratings given by groups of students was commonly high (Marsh and Roche 1997; Richardson, 2005); although the research by Feistauera and Richterb (2017) suggested that student evaluations of teaching can be reliable assessments of the course and the teacher when aggregated evaluations based on a sufficient number of students are used; however, the inter-rater reliability of student evaluations of teaching varied between different measures and course types (seminar and lecture).

The present study was carried out in a different context to most of the previous works (Clayson, 2009). Concretely, it measures the student evaluation of teaching in a higher education institution, the National Polytechnic School of Ecuador, where students are enrolled in technical subjects such as engineering, architecture and biotechnology. Although there were no records in the beginning of teacher evaluations in higher education in Ecuador, this has been a widespread practice in Ecuadorian higher education institutions since the early 1980s (Pareja, 1986).

The student evaluation of teaching instrument used in the National Polytechnic School is the "Cuestionario de Evaluación de la Enseñanza del Profesor de la Escuela Politécnica Nacional del Ecuador" [*Teacher Evaluation Questionnaire of the National Polytechnic School, TEQNS*]. The evolution of the questionnaire consisted of the proposal of several effective teaching criteria, from which a set of items were developed by a teaching committee, which was part of the management team of the *National Polytechnic School.* The aspects to be evaluated and the specific items that make up the questionnaire are approved each academic year by the management team of the National Polytechnic School.

To determine the dimensions of teaching competence and more specific aspects to evaluate, both conceptual and applied aspects were taken into account. Based on these two approaches, Apodaca and Grad (2005) proposed five dimensions: 1. planning and preparation, 2. communication skills 3. interaction with students 4. didactics and methodological resources, and 5. assessment.

From the perspective of academic staff training, Newble and Cannon (1995) proposed 'organization', 'instruction', 'evaluation', 'relationships' and 'subject mastery' as the most important aspects of teaching. Therefore, the items are grouped theoretically into the following four factors: 1. *Planning, mastery and clarity in the explanation of the subject matter*, that incorporates the knowledge of subject, clarity and understanding, including sensitivity to and concern with class level and progress, structure, planning, preparation and organization of the course (i.e. *The teacher appropriately*

selected the class activities, according to the objectives). 2. *Methodology and resources,* referring to the use of appropriate and varied teaching methods and materials (i.e. *The teacher conveniently used different teaching methods*). 3. *Evaluation,* understood as the use of objective and impartial methods, related to teaching, and useful to reorient student learning (i.e. *The teacher evaluated fairly and impartially*). 4. *Teacher-student relationship*, referring to concern and respect for students, friendliness of the teacher, rapport, openness to opinions of students, encouragement of the student initiatives, availability, and helpfulness (i.e. *The teacher has been given suggestions that he/she accepted openly*).

Although the number and dimensions of effective teaching have remained an open question (Spooren et al., 2013), these four dimension are present in the most of SET rating scales literature, and they are aspects related to the teaching effectiveness (Apodaca and Grad, 2005; Cohen, 1981; Feldman, 1989; Huybers, 2014; Richardson, 2005; Spooren, et al., 2013).

Content and face validity are taken mainly into account in the development of the questionnaires. However, the empirical validation is minimal and is limited to the descriptive and discriminatory analysis of the items individually considered. It is lacking a complete process of construct and criterion validity, just as the estimation of the reliability of the scale and/or the subscales that make up these instruments. In this regard, item response theory (IRT) has the main advantage of focusing on the quality of items in measuring underlying constructs (Van der Linden, 2017). IRT models give researchers more confidence in applying the scale in wider contexts (Wright & Masters, 1982). Differing from classical test theory, which considers that an observed test score is composed by a true score and a random error component, IRT considers that the probability of a person's expected response to an item is a mathematical function of that person's ability and one or more parameters characterizing the item (Bond & Fox, 2015, p. 363).

The Rasch model (Rasch, 1960) is the most well-known among unidimensional item response theory (UIRT) models, providing a method based on the calibration of ordinal data from a shared measurement scale and enabling one to test conditions such as dimensionality, linearity and local independence. Calibration is the procedure used to estimate personal latent traits or item difficulty by converting raw score odds to logits on an IRT measurement scale (Bond & Fox, 2015). Moreover, as the TEQNS has different factors, a multidimensional item response theory (MIRT) is a better technique to simultaneously calibrate all subscales and increase the measurement precision by taking into account the correlation between subscales (Adams, Wilson, & Wang, 1997).

The aim of the present study was to examine the internal validity of the TEQNS, applying both unidimensional item response theory (UIRT) and multidimensional item response theory (MIRT) models to examine and compare different theoretical internal structures of the instrument. Concretely, three research questions were considered:

Research question 1 (RQ1): Do all the TEQNS items represent a unique construct?

Research question 2 (RQ2): Do each of the different factors of the TEQNS represent a single unidimensional construct?

Research question 3 (RQ3): Are the factors of the TEQNS interrelated?

2 Method

2.1 Participants

The sample consisted of 6110 students of the National Polytechnic School from Ecuador, who rated the teaching of their 310 teachers, who composed a varied sample of age, category, and teaching experience. These students were enrolled in 8 different Faculties and Schools, in 28 different degree programs, and attended 358 different classes. In the population, 68.3% of the students were male and 31.7% female. The distribution of students and percentages per academic department and undergraduate degrees can be seen in Table 1. The higher percentage of male students was representative of the population of students of polytechnic studies. The average age was 22.6 years old (SD = 3.2). These students rated the faculty's teaching during the 2016/17 academic year.

Table 1 Frequencies and gender distribution from the TEQNS dataset by faculty and undergraduate degree

Faculty	Undergraduate degree	Frequency	Gender	
			M	F
Basic Sciences	Mathematics	33	20	13
	Physics	106	80	26
	Basic Sciences	90	62	28
Administration Sciences	Bussiness Studies	315	180	135
	Economic and Finantial Studies	232	150	82
	Social Sciences	282	145	137
Civil and Environmental Engineering	Civil Engineering	342	222	120
	Environmental Engineering	264	166	98
	Water Technology	168	88	80
Electronic Engineering	Electric Engineering	281	192	89
	Electronic and control Engineering	462	351	111
	Information network Engineering	225	157	68
	Telecommunitations Engineering	359	244	115
	Mathematical Engineering	397	313	84
	Electromechanical Technology	172	113	59
	Electronical Technology and Telecommunications	350	255	95
Geology and Petroleum	Geological Engineering	111	81	30
	Petroleum Engineering	202	177	25
Mechanical Engineering	Mechanical Engineering	710	532	178
Chemical engineering and Agribusiness	Chemical Engineering	283	187	96
	Agribunsiness Engineering	170	102	68
Computer Systems Engineering	Computer Systems and Computing Engineering	439	285	154
	Computer Technology	117	71	46
Total		6110	4173	1937

Note. M= Male. F= Female

2.2 Measures

Students' evaluations of teaching ratings were obtained from the "Cuestionario de Evaluación de la Enseñanza del Profesor de la Escuela Politécnica Nacional del Ecuador", [*Teacher Evaluation Questionnaire of the National Polytechnic School*], approved by the teaching staff for the 2016-17 academic year.

The items are grouped theoretically into four factors: 1. Planning, mastery and clarity in the explanation of the subject-matter (items 1-9), (i.e.: *The teacher conveniently expressed the objectives and themes, indicating their relationship with the professional training of the studies taken*). 2. Methodology and resources (items 10-15), (i.e.: *The teacher explained didactic material apart from the textbook and made it understandable*). 3. Evaluation (items 16-23) (i.e.: *The evaluation events are related to the teaching given*). 4. Teacher-student relationship (items 24-32 items) (i.e.: *The teacher created a climate of trust and work in class*). Students responded to these items on a 5-point rating scale ranging from 1 (do not agree at all) to 5 (totally agree). The full scale, with the items grouped into the four theoretical dimensions are included in the Annex, in original Spanish version and English translation.

2.3 Procedure

The data collection was made from the existing computer records in the administration of the National Polytechnic School of the Ecuador and, access was granted with permission of the Vice Chancellor for Academic Affairs of the Institution. The data provided by the institution were anonymous, with only one identification code for each student. Students' age, and gender, as well as teachers' age, gender, and experience were collected from administrative records.

The application of the scale of evaluation of teaching by the student was carried out towards the end of the semester, before they knew their final grades. All the teachers were evaluated by the students in a similar period of time. All the students had to evaluate the teachers to be able to access their final grades. The student evaluation of teaching was made through an electronic platform, in which the data were recorded.

The impact of faculty procedures on response rates of student evaluations of teaching has been studied by several authors, as opposed to special electronic evaluations. So, Young, Joines, Standish, and Gallagher (2019) found that response rates were substantially higher when faculty provided in-class time for students to complete student evaluations of teaching compared to the electronic form from the administration. However, there are studies designed to analyze this question that do not find differences between the evaluations with electronic questionnaires and those with paper and pencil; additionally, this is true when a more representative sample responds instead of a smaller and biased sample (Nowell, Gale & Kerkvliet, 2014).

Once, the response rate in electronic administration was lower than with paper-and-pencil questionnaires, so the procedure followed in this case consisted of forcing all the students to answer the evaluation survey in order to access their final grades. This procedure has proved useful and valid in some higher education institutions (Leung, & Kember, 2005; Nair, & Adams, 2009).

2.4 Data analysis

Two indicators, item weighted fit and "expected a posteriori/plausible value (EAP/PV)" reliability, were examined. Item weighted fit was selected as it indicates how well the item parameters of a measure fit the empirical dataset. Concretely, infit and outfit statistics were used to check the quality of the instrument. These indexes are the mean value of the squared residuals. Therefore, the larger the squared residual, the larger misfit between data and model.

The infit statistic is an information-weighted sum, so this variance is larger for well-targeted observations and smaller for extreme observations (Bond & Fox, 2015). d values of outfit and infit mean squares can range from 0 to positive infinitive. Values below 1 indicate a higher than expected fit of the model, whereas values greater than 1 indicate poor fit of the model. An infit/outfit range between 0.75 and 1.33 can be considered as acceptable values (Wu, Adams, & Wilson, 1998).

With respect to EAP/PV reliability, it is the ratio of modeled variance to observe variance. The interpretation of this parameter is similar than Cronbach's alpha, such that a reliability .70 is considered to be the minimum standard, while .80 is recommended for screening purposes (Salvia, Ysseldyke, & Bolt, 2013).

The analyses were involved in three phases: In the first phase, the UIRT of model 1 was examined, which is based on the assumption that all the items load on a unique construct, namely "student evaluation of teaching". Item weighted fit and EAP/PV reliability, and principal component analyses were the indicators used to examine this model. Winsteps version 4.5 statistical software (Linacre, 2019b) was used to check whether the items in each subscale satisfy the unidimensionality assumption. Unidimensionality requires that the measurement should target one attribute or dimension at one time (Bond & Fox, 2015). According to Linacre (2019a), an eigenvalue less than 2.0 of the first contrast indicated the residuals are not relevant enough to disturb the measurement quality. An eigenvalue more than 2.0 implies that there is probably another dimension in the measurement instrument.

In the second phase, it examined the quality measurement of the items within each domain independently, which a total of 4 factors (e.g., Model 2-PMC; Model 2-MR; Model 2-E, and Model 2-TR). Item weighted fit and EAP/PV reliability were the indicators used to examine the goodness of fit.

The third phase consisted of examining Model 3, which represented the hypothetical model constructed for the TEQNS. To evaluate it, Model 3 was calibrated to evaluate the item weighted fit, EAP/PV reliability, and the correlation between latent traits by domain. Results from Model 3 were first compared with results from Model 2 to see whether the EAP/PV reliability for each factor was improved. Results from Model 1 were also compared with Model 3 in order to assess which model shows the best underlying measurement description of both items and factors. To this end, deviance (-2log likelihood) and Akaike information criterion (AIC) were employed (Wu, Adams, Wilson, & Haldane, 2007).

The internal structure of the TEQNS was tested through various IRT models using Conquest version 2.0 (Wu et al., 2007). Each of the three models are summarized in Table 2. All models are based on the rationale of the Rasch Rating Scale Model (RRSM). Maximum likelihood estimation method

was employed for the parameters of the model. The Monte Carlo method was used for calibration in all the models to ensure index comparison.

3 Results

With respect to the first model, the TEQNS was considered to measure a unidimensional construct, namely, student evaluation of teaching as an overall composed dimension. The weighted fit was acceptable in all items with the exception of item 11 (see Table 3). The EAP/PV reliability was 0.91, considered as an excellent value. For this reason, a complement analysis of unidimensionality was used. A principal component analysis of the residual scores (Linacre, 1998; Wright, 1996) showed that the eigenvalue of the first contrast for the whole scale was 3.61; the second contrast was 2.64 and the third contrast was 2.48. Therefore, the instrument could be considered as multidimensional.

Table 2 Summary of IRT models

	Model 1	Model 2	Model 3
IRT model	UIRT	UIRT	Between-item MIRT RRSM
Theory	All items represent a single construct	Items within each factor represent independent constructs	All items represent four interrelated domains.
Number of items	32	F1= 9 items; F2 = 6 items; F3 = 8 items; F4 = 9 items	32 items (9 in F1; 6 in F2; 8 in F3; 9 in F4).
Answering research question	1	2	3

Note. Model 2 represented each of the four domains, named as Model 2-PMC (Planning, mastery and clarity in the explanation of the subject), Model 2-MR (Methodology and resources), Model 2-E (Evaluation), and Model 2-TR (Teacher-student relationship). IRT = Item Response Theory; UIRT = unidimensional item response theory; RRSM = Rasch Rating Scale Model; MIRT = multidimensional item response theory.

Table 3 Fit statistics information for models

	Weighted fit				Misfit items	
Model	Minimum	Maximum	M	SD	Overfit	Underfit
Model 1	0.81	1.18	1.00	0.18	item 11	-
Model 2-PMC	1.00	1.34	1,13	0,11	item 8	-
Model 2-MR	0.92	1.44	1.06	0.20	item 11	-
Model 2-E	0.93	1.45	1.08	0.18	item 16	-
Model 2-TR	0.94	1.30	1.08	0.11		-
Model 3	0.85	1.68	0.99	0.15	item 11	-

Note. "-" indicates that no underfit values were found; PMC = Planning, mastery and clarity in the explanation of the subject; MR = Methodology and resources; E = Evaluation; TR = Teacher-student relationship

In Model 2, the four factors of the TEQNS were calibrated independently within four unidimensional structures. Results indicated that only Model 2-TR had adequate fit statistics, whereas both in Model 2-PMC and Model 2-MR had one misfit item each. In all the models, the EAP/PV values were adequate.

Model 3 was calibrated within MIRT, in which the four domains were inter-correlated. Results from the weighted fit statistics showed adequate model fit for most of the items, as presented in Table 3. Similar to Model 1 and Model 2-MR, item 11 presented extreme fit values. EAP/PV reliabilities showed excellent values, all of the higher than .90 as can be seen in Table 4. Correlations between domains are showed in Table 5. As can be observed, all the correlations ranges were high, and even the correlation between Evaluation and Teacher-student relationship was extremely high (r = .93). When comparing Models 1 and Model 3, it seems that the last set showed better deviance and AIC values (see Table 6). The deviance difference can be considered as statistically significant.

Given that Model 3 showed the best psychometric properties, item 11 was removed due to its misfit value, which was consistent in Model 1 and Model 2-MR. A re-analysis of Model 3 without this item showed appropriate weighted fit values for all the items, ranging from 0.91 to 1.18.

Table 4 EAP/PV Reliability Summarized by Models.

Model	PMC	MR	E	TR	Total
Model 1	-	-	-	-	0.91
Model 2-PMC	0.89	-	-	-	
Model 2-MR	-	0.91			
Model 2-E	-	-	0.99	-	-
Model 2-TR	-	-	-	0.91	-
Model 3	0.95	0.94	0.98	0.95	-

Note. "-"indicates the EAP/PV reliability was not calculated by the specific model. EAP/PV = expected a posteriori/plausible value; PMC = Planning, mastery and clarity in the explanation of the subject; MR = Methodology and resources; E = Evaluation; TR = Teacher-student relationship.

Table 5 Correlation between subdomains estimated for Model 3

Domains	PCM	MR	E
PCM			
MR	0.87		
E	0.90	0.90	
TR	0.87	0.88	0.93

Note. PCM = Planning, mastery and clarity in the explanation of the subject; MR = Methodology and resources; E = Evaluation; TR = Teacher-student relationship.

Table 6 Comparison of Model fit statistics

Model/Comparison		Deviance	AIC	N of parameters
Model				
	Model 1 (unidimensional)	125004.41	125076.41	36
	Model 3(multidimensional)	118823.56	118913.56	45
Model Comparison				
	Model 1-Model 3	6180.85*	6162.85	9

Note. AIC = Akaike information criterion.

*p<.001 when deviance is greater than the critical value of chi-square distribution (χ^2 = 27.87, df = 9).

4 Discussion

Student evaluation of teaching is considered as one of the most important procedures in different kinds of institutions, which include a variety of rating procedures with a lack of a clear consensus (Hornstein, 2017). Based on an extension of the actiotope model of giftedness, which considers the relevance of educational and learning capitals on achievement, this study applied both UIRT and MIRT models to analyze the underlying structure of the TEQNS on a large sample of university students from the National Polytechnic School of Ecuador. This instrument was developed to address methodological problems that exist in the current teacher- student evaluation measures, mainly related to construct validity.

In the first place, in response to RQ1, a calibration of unidimensional measure of SET, showed adequate fit statistics in most of the items, as well as a high total reliability score. This suggested that the total score of the TEQNS may reflect a composite score of the construct. The results also addressed the possibility that the instrument could measure four unidimensional and independent domains (RQ2). In this case, the results showed a higher number of misfit items, and adequate reliability values. In addition, the third question (RQ3) was based on the construction of four inter-correlated domains through a MIRT model, which was supported by adequate reliability and correlation between domains.

When comparing the deviance and AIC of Model 3 with the deviance and AIC of Model 1, it showed that Model 3 has a better description of the data. This indicated that the nature of the internal structure in TES was multidimensional and involved multiple distinct domains. Furthermore, when comparing Model 3 and Model 2, it is recognized that Model 3 has better EAP/PV reliability estimates in three out of the four domains.

It is important to mention that item 11 (*The teacher organized didactic experiences such as visits, excursions, projects, discussions*) showed a constant misfit in all the models where it was included (Model 1, Model 2-PMC, and Model 3). This is a concern affecting the underlying construct. It is possible that the content of this item may not have a direct relation with the student's conception of methodological and didactic resources, but with emotional or social group experiences that exceeds the teach-

ers' organizational features. For this reason, a new calibration of Model 3 without this item was considered, detecting a better distribution of item parameters, with weighted fit statistics rating from 0.91 to 1.18 and no misfit items in any dimension.

This is the first study that implements MIRT model as an effective measure of SET in higher education. Furthermore, given the need to have valid and reliable tools for the assessment of SET in Latin America, the construct validation of the TEQNS has a strong role to play in the assessment of teachers in large public universities.

The current study provides a dynamic perspective on domains of SET, adding important information regarding not only general model fit indexes (as happens with traditional Exploratory or Confirmatory Factor Analysis), but also specific parameters which measure the adequacy of each item as an effective indicator of the underlying construct. IRT can be considered as an effective modeling approach to ensure SET valid measures in the dynamic perspective of educational and learning capitals (Ziegler & Baker, 2013; Ziegler et al., 2017). In line with the characterizing features of a synthetic research strategy, the comparison of endorsability of each item is no longer dependent on a specific sample, and the comparison of the student's assessment of teaching is no longer dependent on specific items (Hambleton & Jones, 1993). At the same time, Rasch models can examine how participants understand different response options and provide additional evidence on the internal construct for SET subscales.

4.1 Limitations and future directions

Although a large dataset was employed for psychometric validation of the TEQNS in one of the most important public institutions in Ecuador, bias analysis was not included in this study. Its results were limited to this institution and this type of technical studies.

A special source of bias in SET studies is response bias, a respondent's consistent answering pattern irrespective of the questions presented. Several studies found both acquiescent responding and extreme responding to be consistent traits within individual students; however, when these were statistically controlled, although their effects were reduced, the relationship remained the same (Huybers, 2014; Richardson, 2012).

As a more general sense, bias is present when a known characteristic of students systematically affects their ratings of teachers. The gender of the students is an example of this possible bias in student evaluation of teaching. Previous studies have found that female students on average tend to give significantly higher SET ratings than their male peers (Badri, Abdulla, Kamali, & Dodeen, 2006; Basow, Phelan, & Capotosto, 2006; Boring, 2015; Centra, & Gaubatz, 2000; Darby, 2006).

Another source of bias is the discipline. If the evaluation of teaching is situational and is affected for academic disciplines, being higher in studies in the field of education and the liberal arts, and not as high in other areas, such as business and engineering (Clayson, 2009) it seems necessary to carry out new studies in areas different from the previous ones, as the technical areas where there are less studies on the subject.

Two steps in the development of this measure should be considered: first, personal and psychological variables should be analyzed, including students' demographic characteristics (e. g. geographical regions, race) to check measurement precision of the instrument. Second, further analyses are required to detect possible causes of SET differences between sample subgroups and efficient vs non-efficient teachers' features associated with SET. Such work would further both the reliability and validity of the current measurement using diverse populations and ecological contexts.

5 References

Abrami, C.P., d'Apollonia, S., & Rosenfield, S. (1997). The dimensionality of student ratings of instruction: what we know and what we do not. In R. E. Perry and J.C. Smart (Eds.), Effective teaching in higher education: Research and practice (pp. 321–367). New York: Agathon Press.

Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficient multinomial logit model. *Applied Psychological Measurement, 21*, 1-23. https://doi.org/10.1177/0146621697211001

Apodaca, P., & Grad, H. (2005). The dimensionality of student ratings of teaching: Integration of uni- and multidimensional models. *Studies in Higher Education*, *30*, 723–748. https://doi.org/10.1080/03075070500340101

Arocena, R., & Sutz, J. (2001). Changing knowledge production and Latin American universities. *Research Policy, 30*(8), 1221-1234.

Badri, M. A., Abdulla, M., Kamali, M. A., & Dodeen, H. (2006). Identifying potential biasing variables in student evaluation of teaching in a newly accredited business program in the UAE. *International Journal of Educational Management, 20*(1), 43-59.

Barnes, D., Engelland, B., Matherne, C., Martin, W., Orgeron, C., & Ring, J.(2008). Developing a psychometrically sound measure of collegiate teaching proficiency. *College Student Journal*, *42*, 199-213.

Barrett, P. (2012). *Gower program help file*. Auckland, New Zealand: Advanced Projectes R&D Ltd.

Basow, S. A., Phelan, J. E., & Capotosto, L. (2006). Gender patterns in college students' choices of their best and worst professors. *Psychology of Women Quarterly*, *30*(1), 25-35.

Boice, R. (1991). New faculty as teachers. *Journal of Higher Education, 62*(2), 150-173.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). New York, NY: Routledge.

Centra, J.A., and Gaubatz, N.B. (2000). Is There Gender Bias in Student Evaluations of Teaching? *The Journal of Higher Education*, *71*(1), 17-33. https://doi.org/10.2307/2649280

Cheung, D. (2000). Evidence of a single second-order factor in student ratings of teaching effectiveness. *Structural Equation Modeling: A Multidisciplinary Journal*, *7*(3), 442-460, DOI: 10.1207/S15328007SEM0703_5

Cho, E. (2016). Making reliability reliable: A systematic approach to reliability coefficients. *Organizational Research Methods*, *19*(4), 651-682. https://doi.org/10.1177/1094428116656239

Clayson, D. E. (2009). Student evaluations of teaching: Are they related to what students learn? A meta-analysis and review of the literature. *Journal of Marketing Education*, *31*, 16–30. https://doi:10.1177/0273475308324086

Consejo de Educación Superior (CES) (2017). Reglamento de Carrera y Escalafón del Profesor e Investigador del Sistema de Educación Superior. [Career and Ladder Regulations of the Professor and Researcher of the Higher Education System]. Retrieved on the 20 April 2019, from: https://procuraduria.utpl.edu.ec/sitios/documentos/NormativasPublicas/Reglamento%20de%20Carrera%20y%20Escalaf%C3%B3n%20del%20Profesor%20e%20Investigador%20del%20Sistema%20de%20Educaci%C3%B3n%20Superior%202018.pdf

Darby, J. A. (2006). Evaluating courses: An examination of the impact of student gender. *Educational Studies, 32*(2), 187-199. https://doi.org/10.1080/03055690600631093

Feistauera, D., & Richterb, T. (2017). How reliable are students' evaluations of teaching quality? A variance components approach. *Assessment & Evaluation in Higher Education*, *42*(8), 1263–1279. https://doi.org/10.1080/02602938.2016.1261083

Glaser, R., Lesgold, A, & Lajoie, S. P. (1988). Toward a cognitive theory for the measurement of achievement. In R. Ronning, J. Glover, J. S., Conoley, & J. C. Wittrock (Eds.). *The influence of cognitive psychology on testing*, vol. 3. Hillsdale, NJ: Erlbaum.

Heller, K. A., Perleth, C., & Lim, T. K. (2005). The Munich model of giftedness designed to identify and promote gifted students. In R. J. Sternberg & J. E. Davidson (Eds.), *Conceptions of giftedness* (2nd ed., pp. 147-170). Cambridge, UK: Cambridge University Press.

Hornstein, H. A. (2017). Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance. *Cogent Education, 4*, 1. http://dx.doi.org/10.1080/2331186X.2017.1304016

Huybers, T. (2014) Student evaluation of teaching: the use of best–worst scaling, *Assessment & Evaluation in Higher Education, 39*(4), 496-513. https://doi.org.10.1080/02602938.2013.851782

Keeley, J., Furr, R.M., & Buskist, W. (2010). Differentiating psychology students' perceptions of teachers using the Teacher Behavior Checklist. *Teaching of Psychology*, *37*, 16-20. https://doi.org.10.1080/00982890342682

Larsson, S. (1986). Learning from experience: teachers' conceptions of changes in their professional practice. *Journal of Curriculum Studies, 19*(1), 37-43.

Leinhardt, G., & Greeno, J. G. (1986). The cognitive skill of teaching. *Journal of Educational Psychology, 78*(2), 75-95.

Leung, D.Y.P., & Kember, D. (2005). Comparability of data gathered from evaluation questionnaires on paper and through the Internet. *Research in Higher Education*, *46*(5), 571-591. https://doi.org/10.1007/s11162-005-3365-3

Linacre, J. M. (1998). Structure in Rasch residuals: Why principal component analysis? *Rasch Measurement Transactions, 12*, 636.

Linacre, J. M. (2019a). *A user's guide to Winsteps & Ministeps Rasch-Model computer programs*. Available at http://www.winsteps.com

Linacre, J. M (2019b). *Winsteps* (version 4.4.7.) [computer software]. Chicago: MESA.

Linse, A.R. (2017). Interpreting and using student rating data: Guidance for faculty serving as administrators and on evaluation committees. *Studies in Educational Evaluation*, *54*, 94-106. https://doi.org/10.1016/j.stueduc.2016.12.004

Marsh, H. W. (1982). SEEQ: a reliable, valid and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology*, *52*, 77-95.

Marsh, H. W. (1987). Student's evaluations of university teaching: Research findings, methodological issues, and directions for further research. *International Journal of Educational Research*, *11*, 253–388. https://doi.org/10.1016/0883-0355(87)90001-2

Marsh, H. W. (1991a). A multidimensional perspective on student's evaluations of teaching effectiveness: Reply to Abrami and d'Apollonia (1991. *Journal of Educational Psychology*, *83*,(3), 416–421. https://doi.org/10.1037/0022-0663.83.2.285

Marsh, H. W. (1991b). Multidimensional students' evaluations of teaching effectiveness: A test of alternative higher-order structures. *Journal of Educational Psychology*, *83*, 285–296. https://doi.org/10.1037/0022-0663.83.2.285

Marsh, H. W. (2001). Distinguishing between good (useful) and bad workloads on students' evaluations of teaching. *American Educational Research Journal*, *38*, 183–212. https://doi.org/10.3102/00028312038001183

Marsh, H. W. (2007a). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence based perspective* (pp. 319–383). New York: Springer.

Marsh, H. W. (2007b). Do university teachers become more effective with experience? A multilevel growth model of students' evaluation of teaching over 13 years. *Journal of Educational Psychology*, *99*, 775–790. https://doi.org/10.1037/0022-0663.99.4.775

Marsh, H.H., Hau, K.T., Balla, J.R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research*, *33*, 181-230.

Marsh, H.W., & Hocevar, D. (1991). Students' evaluations of teaching effectiveness: the stability of mean ratings of the same teachers over a 13-year-period. *Teaching and Teacher Education*, *7*, 303-314.

Marsh, H. W., Martin, A. J., & Jackson, S. E. (2010). Introducing a short version of the physical self description questionnaire: new strategies, short-form evaluative criteria, and applications of factor analyses. *Journal of sport & exercise psychology*, *32*(4), 438-482. https://doi.org/10.1123/jsep.32.4.438

Marsh, H. W., Muthèn, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling*, *16*, 439–476. https://doi.org/10.1080/10705510903008220

McNeish, D. M. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*(3), 412-433. https://doi.org/10.1037/met0000144

Mezirow, J. (1991). *Transformative dimensions of adult learning*. San Francisco: Jossey-Bass.

Morin, A. J. S., Katrin Arens, A., & Marsh, H. W. (2016). A bifactor exploratory structural equation modeling framework for the identification of distinct sources of construct-relevant psychometric multidimensionality. *Structural Equation Modeling*, *23*(1), 116-139. https://doi.org/10.1080/10705511.2014.961800

Mortelmans, D., & Spooren, P. (2009). A revalidation of the SET37-questionnaire for student evaluations of teaching. *Educational Studies*, *35*, 547-552. https://doi.org/10.1080/03055690902880299

Muthen, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology*, *25*, 267-316. https://doi.org/10.2307/271070

Muthén, L. K., & Muthén, B. O. (2018). *Mplus user's guide* (8.ª ed.). Los Angeles, CA: Muthén & Muthén.

Nair, C.S., & Adams, P. (2009) Survey Platform: A Factor Influencing Online Survey Delivery and Response Rate, *Quality in Higher Education*, *15*(3), 291-296. https://doi.org/10.1080/13538320903399091

Nagy, M. S. (2002). Using a single-item approach to measure facet job satisfaction. *Journal of Occupational and Organizational Psychology*, *75*(1), 77–86. http://dx.doi.org/10.1348/096317902167658

Newble, D. & Cannon, R. (1995). *A handbook for teachers in universities & colleges: a guide to improving teaching methods*. London: Kogan Page.

Nowell, C., Gale, L.R., Kerkvliet, J. (2014). Non-response bias in student evaluations of teaching. *International Review of Economic Education*, *17*, 30-38.

Pajares, M. F. (1992). Teachers' beliefs and educational research: cleaning up a messy construct. *Review of Educational Research, 307*-332.

Pareja, F. (1986). *La educación superior en el Ecuador* [The higher education in Ecuador]. Caracas: Regional Center for Higher Education in Latin America And the Caribbean (CRESALC)-UNESCO.

Pratt, D. D. (1992). Conceptions of teaching. *Adult Education Quarterly, 42*(4), 203-220.

Ribeiro, D. (1971). *La Universidad Latinoamericana [The Latin-American University]*. Caracas: Ediciones de la Biblioteca de la Universidad Central de Venezuela.

Richardson, J.T.E. (2005). Instruments for obtaining student feedback: a review of the literature. *Assessment & Evaluation in Higher Education*, *30*:4, 387-415. http://dx.doi.org/10.1080/02602930500099193

Richardson, J.T.E. (2012). The role of response bias in the relationship between students' perceptions of their courses and their approaches to studying in higher education. *British Educational Research Journal*, *38*, 399-418. https://doi.org/10.1080/01411926.2010.548857

Salvia, J., Ysseldyke, J. E., & Bolt, S. (2013). *Assessment in special and inclusive education* (12th ed.). Belmont, CA: Wadsworth.

Saroyan, A., & Amundsen, C. (2001). Evaluating university teaching: Time to take stock. *Assessment & Evaluation in Higher Education, 26*(4), 341-353. https://doi.org/10.1080/02602930120063493

Scherer R, Nilsen T., & Jansen M. (2016). Evaluating individual students' perceptions of instructional quality: An investigation of their factor structure, measurement invariance, and relations to educational outcomes. *Frontiers in Psychology*, *7*:110. https://doi.org/10.3389/fpsyg.2016.00110

Schmitt, T. A., Sass, D. A., Chappelle, W., & Thompson, W. (2018). Selecting the "best" factor structure and moving measurement validation forward: An illustration. *Journal of Personality Assessment*, *100*(4), 345-362. https://doi.org/10.1080/00223891.2018.1449116

Schön, D. A. (1983). *The reflective practitioner: how professionals think in action*. San Francisco: Jossey-Bass.

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the Validity of Student Evaluation of Teaching: The State of the Art. *Review of Educational Research XX(X), pp. 1–45***.** http://dx.doi.org/10.3102/0034654313496870

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics.* Boston: Pearson.

Toland, M., & De Ayala, R.J. (2005). A multilevel factor analysis of students' evaluations of teaching. *Educational and Psychological Measurement*, *65*, 272-296. https://doi.org/10.1177/001316440426866

Uttl, B., White, C.A., & Gonzalez, D.W. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation*, *54*, 22-42. https://doi.org/10.1016/j.stueduc.2016.08.007

Van der Linden, W. J. (2017). *Handbook of Item Response Theory, Volume Three: Applications.* New York, NY: Chapman and Hall/CRC.

Vandenberg, R. J., & Lance, C. E. (2000). A Review and synthesis of the measurement invariance literature: Suggestions , practices , and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4-70. https://doi.org/10.1177/109442810031002

Veas, A., Castejón, J. L., O'Reilly, C., & Ziegler, A. (2018). Mediation Analysis of the Relationship between Educational Capital, Learning Capital, and Underachievement among gifted secondary school students. Journal for the Education o the Gifted, 41(4), 369-385. https://doi.org/10.1177/0162353218799436

Wright, B. D. (1996). Local dependency, correlations and principal components. *Rasch Measurement Transactions, 10*(3), 509-511.

Wu, : L., Adams, R. J., & Wilson, M. R. (1998). ACER Conquest: Generalised item response modelling software [Computer software and manual]

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. (2007). *ACER Conquest version 2.0: Generalised item response modelling software*. Camberwell, Victoria: Australian Council for Educational Research.

Xiao, Y., Liu, H., & Hau, K.-T. (2019). A comparison of CFA, ESEM, and BSEM in test structure analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, In press. https://doi.org/10.1080/10705511.2018.1562928

Young, K., Joines, J., Standish, T., & Gallagher, V. (2019) Student evaluations of teaching: the impact of faculty procedures on response rates, *Assessment & Evaluation in Higher Education*, *44* (1), 37-49, https://doi.org/10.1080/02602938.2018.1467878

Zabaleta, F. (2007). The use and misuse of student evaluation of teaching. *Teaching in Higher Education*, *12*, 55–76. https://doi.org/10.1080/13562510601102131

Ziegler, A., & Baker, J. (2013). Talent development as adaption: The role of educational and learning capital. In S. Phillipson, H. Stoeger, & A. Ziegler (Eds), *Exceptionality in East-Asia: Explorations in the actiotope model of giftedness* (pp. 18-39). Routledge.

Ziegler, A., Chandler, K., Vialle, W., & Stoeger, H. (2017). Exogenous and endogenous learning resources in the Actiotope Model of Giftedness and its significance for gifted education. *Journal for the Education of the Gifted, 40*, 310-333. https://doi.org/10.1177/0162353217734376

Ziegler, A., Debatin, T., & Stoeger, H. (2019). Learning resources and talent development from a systemic point of view. *Annals of the New York Academy of Sciences*, 1445, 39—51. https://doi.org/10.1111/nyas.14018

Ziegler, A., & Stoeger, H. (2017). Systemic gifted education. A theoretical introduction. *Gifted Child Quarterly, 61*, 183–193. doi:10.1177/0016986217705 713

Ziegler, A., Stoeger, H., & Balestrini, D. (2017). Systemic gifted education. In J. Riedl Cross, C. O'Rellly & T. Cross (Eds.), *Providing for the special needs of students with gifts and talents* (pp. 15-55). Dublin, Ireland: Kazoo Independent Publishing.

Corresponding author:
Alejandro Veas, PhD
University of Alicante.
Department of Developmental Psychology and Didactics
PO 03690
San Vicente del Raspeig, Alicante, Spain
E-mail: Alejandro.veas@ua.es