Assessing the Students' Evaluations of Educational Quality (SEEQ) questionnaire in Greek higher education

Vasilis Grammatikopoulos, M. Linardakis, A. Gregoriadis & V. Oikonomidis

Higher Education
The International Journal of Higher Education Research

ISSN 0018-1560
Volume 70
Number 3

High Educ (2015) 70:395-408
DOI 10.1007/s10734-014-9837-7

Springer

High Educ (2015) 70:395–408
DOI 10.1007/s10734-014-9837-7

Assessing the Students' Evaluations of Educational Quality (SEEQ) questionnaire in Greek higher education

Vasilis Grammatikopoulos · M. Linardakis · A. Gregoriadis · V. Oikonomidis

Published online: 7 December 2014

Abstract The aim of the current study was to provide a valid and reliable instrument for the evaluation of the teaching effectiveness in the Greek higher education system. Other objectives of the study were (a) the examination of the dimensionality and the higher-order structure of the Greek version of Students' Evaluation of Educational Quality (SEEQ) questionnaire, and (b) the investigation of the effects of several background variables on students' evaluations of teaching (SET) scores provided by the Greek version of SEEQ. A total of 1,264 students participated by filling in the questionnaires administered to them. The participants were selected from social science departments that belonged to eight universities of Greece. The results showed solid evidence of the applicability of the Greek version of SEEQ, by confirming the factor structure of the instrument and reassuring the multidimensionality of the teaching effectiveness construct. Additionally, the effects of several background variables on teaching effectiveness further supported the validity of SET scores.

Keywords Educational evaluation · Students' evaluations of teaching (SET) · Teaching effectiveness · SEEQ

Introduction

The students' evaluations of teaching effectiveness (SET) in higher education

Students' evaluations of teaching (SET) are universal measures and have been applied in almost every higher educational system in the world (Zabaleta 2007). Universities have

V. Grammatikopoulos (✉) · M. Linardakis · V. Oikonomidis
School of Education, University of Crete, University Campus, 74100 Rethymnon, Crete, Greece
e-mail: gramvas@uoc.gr

A. Gregoriadis
School of Education, Aristotle University of Thessaloniki, School of Education Building, 54124 Thessaloniki, Greece

been developing and implementing SET instruments, analysing the results, and using them for the evaluation of the provided instructional quality (Spooren et al. 2013).

The expansion of SET effectiveness in higher education during the previous decades has triggered an ongoing discussion about their validity. Aleamoni (1987, 1999) supported the validity of SET with his work 'The Student Rating Myths versus Research Facts' in which he presented common 'myths' about SET and claimed that faculties and administrators manufactured them in order to argue against the value of SET; on the one hand, studies challenge the validity of SET (Feely 2002; Germain and Scandura 2005; Safer et al. 2005). On the other hand, a plethora of well-designed studies confirmed the usefulness and validity of SET (Alsmadi 2005; Balam and Shannon 2010; Coffey and Gibbs 2001; Marsh 2007a, b).

There are also some recent review studies that did not offer support to either of the two opposite opinions (Jones et al. 2014; Spooren et al. 2013). For example, in their extended literature review on the validity of SET, Spooren et al. (2013) could not draw firm conclusions. The researchers argued that SET validity is possibly affected by the variety of methods, measures and populations used in the studies and pointed that in most cases, institutional designed instruments were used instead of widely accepted standardized instruments (e.g. Student Evaluation of Educational Quality, Course Experience Questionnaire). Another recent study investigated the effect of students' personality traits on SET scores, finding a clear and consistent relationship between them (McCann and Gardner 2014), whereas Boysen et al. (2014) revealed the possible misinterpretation of SET in terms of generalization from limited data. Spooren et al. (2013) presented a very detailed overview of possible student, teacher and course-related characteristics that might affect SET scores. They indicated that not all characteristics could be considered as biasing factors, since some of them (e.g. student effort, class attendance, Prior Subject Interest) are indicators of learning and are reasonably related to SET scores.

In addition to the existing arguments questioning the validity of SET, many researchers (e.g. El Hassan 2009; Marsh 1982a, 1987, 2007a; Spooren et al. 2013) argued that in most cases, students actually evaluate the instructor of the course and not the instructional approach or the quality of learning. Thus, SET data have to be handled with caution and faculties also have to rely on other sources of information (d' Apollonia and Abrami 1997; Marsh 2007a), especially if SET scores are going to be used for employment decisions (Jones et al. 2014). This contradictory context is furthermore empowered by research evidence supporting that there is not a single criterion for teaching effectiveness (Abrami et al. 1990; Abrami and Mizener 1983; Marsh 1987, 1994, 1995; Marsh and Roche 1997), and thus, the validity question is difficult to be universally answered.

The criticism on the validity of SET scores led to the inquiry of alternative sources for the evaluation of teaching effectiveness. Such sources might be committees of experts, peer evaluation, self-evaluation, students' grades, etc. Yet, research evidence revealed that SET scores were more valid and reliable than any other source, and thus, data from other sources shall be used together with SET scores and not instead of them (Cashin 1989, 1995; Chism 1999; Marsh 1987; McKeachie 1997; Zhao and Gallant 2012).

The development of new instruments by each university might not be an ideal solution, because instruments that are applied only in one setting are not easily 'evaluated in relation to rigorous psychometric considerations and revised accordingly' (Marsh 2001, p. 4). The use of standardized and widely implemented instruments can result to valid and reliable SET scores. A lot of SET instruments have been developed and applied over the past few decades (see review of Spooren et al. 2013).

The instrument used in the current study is the Students' Evaluations of Educational Quality (SEEQ). It is a questionnaire developed about 30 years ago (Marsh 1982b), and it is considered one of the most widely used and universally accepted instruments. SEEQ has successfully provided valid and reliable SET scores in a variety of environments, such as several different higher education settings and different countries (e.g. Australia, USA, UK, Hong Kong, China, Spain, India) (Balam and Shannon 2010; Coffey and Gibbs 2001; Marsh 1986; Marsh et al. 1997; Watkins and Thomas 1991).

Apart from its international recognition, another critical point for choosing SEEQ as the instrument of this study was the theoretical basis on which it was developed. Surprisingly, a lot of other existing SET instruments did not take under consideration the theories of teaching and learning in higher and adult education. Marsh and Dunkin (1992) evaluated the content of SEEQ in relation to general principles of teaching and learning in post-secondary education reported by Feldman (1976) and Fincher (1985). They revealed that SEEQ factors adequately included the principles described on the aforementioned studies. Moreover, the 'superiority' of SEEQ against other SET instruments also relies on psychometric analyses, as it constantly reveals high levels of validation and reliability scores (Coffey and Gibbs 2001; Marsh 1987; Marsh and Hocevar 1991a).

Evaluation of teaching effectiveness in the Greek higher educational setting

Until 2008, the Greek higher education had not established any official evaluation procedures for teaching effectiveness. Despite the fact that academic staff has to be evaluated for teaching effectiveness in order to be promoted in a higher rank or get appointed, these procedures were not officially established. A possible explanation is that in Greece, all academic promotions are open to any candidate, something that hampers the assessment of the teaching effectiveness of academics applying for the same position and coming from different universities, different countries or even different departments.

In 2005, the Hellenic Quality Assurance and Accreditation Agency for Higher Education (http://www.adip.gr) was founded, and one of the duties assigned to the Agency was the evaluation of teaching effectiveness in the Greek higher education.

The Agency adopted an internal evaluation procedure in 2007, and SET questionnaires were developed in order to assess the teaching effectiveness of the faculty staff. Yet, instead of choosing a standardized instrument, the Agency proposed some general factors and items and then let the university departments to freely adjust these instruments by adding, changing or removing items. This variety of SET instruments limited the possibility of comparing results and drawing overall conclusions. Marsh (2001) revealed that 'home-made' instruments are unlikely to evaluate the dimensions of teaching effectiveness broadly, and their usefulness can be questioned.

Multidimensionality and higher-order structure of SET

SEEQ is an instrument based on the perception that teaching effectiveness is a multidimensional construct and that SET scores have to be interpreted as such. This perception has been demonstrated in many studies conducted by Marsh (1982b, 1984, 1987, 1991a, b) and colleagues (Marsh and Dunkin 1992; Marsh and Hocevar 1991a). However, there are also studies arguing against the multidimensionality interpretation of SET scores (Abrami and d'Apollonia 1991; Apodaca and Grad 2005). Apodaca and Grad (2005) claimed that SET scores could be treated as multidimensional as much as unidimensional, by revealing that the scores in their study showed multidimensional structure, yet with the presence of a

general factor capturing the overall teaching effectiveness. Abrami and d'Apollonia (1991) claimed that the dimensions of SET scores could be subsumed by a single overarching construct based either on overall items or on a weighted average of items. On the other hand, Marsh argued that SET scores could not be adequately captured by a single higher-order factor. This debate flourished because SET is used not only for feedback, but also for administrative purposes, where a single score representing an overall assessment of the instructional competence is very desirable.

Marsh (1991b) supported the higher-order structure of the teaching effectiveness notion, but also indicated that a unidimensional interpretation of SET scores could be possible, only by weighting the multiple dimensions of teaching effectiveness differently. Recent studies also provided support for the higher-order structure of teaching effectiveness (Apodaca and Grad 2005; Burdsal and Harrison 2008; Cheung 2000; Harrison et al. 2004; Mortelmans and Spooren 2009). In their review study, Spooren et al. (2013) argued that SET instruments should cover the multidimensionality of the notion. They added that scores should be multidimensionally interpreted for feedback and formative purposes, but when it comes to unidimensionally interpretation, they had better be weighed.

The effect of background variables on SET

A very frequently asked question has to do with the possible effect that several background variables might have on SET scores, the 'bias question' (Spooren et al. 2013). There are variables that cannot be considered as 'bias factors' as they are meaningful criteria of teaching effectiveness (e.g. class attendance or student effort), and their relation to SET scores implies valid teaching effectiveness effects rather than bias (Marsh and Roche 1997, 2000; Spooren et al. 2013). Research interest is focused on whether factors that are not directly related to teaching effectiveness (e.g. instructor's gender or course discipline) affect SET scores (Centra and Gaubatz 2000).

The discussion among researchers focuses mainly on whether some factors can be considered as bias (leniency hypothesis) or as meaningful dimensions (validity hypothesis) (e.g. the Expected Grade or Workload/Difficulty factors) (Spooren et al. 2013). Leniency hypothesis claims that instructors can 'buy' favourable evaluations by giving higher grades or reducing the workload and difficulty of a course (see, e.g. Langbein 2008; McPherson and Jewell 2007). On the other hand, many studies argued for the validity hypothesis, proposing that positive relationships, for example between SET scores and Expected Grade are generated by the quality of learning that occurred during the course, and therefore, higher grades should be expected (Marsh and Roche 2000). Moreover, Marsh and Roche (2000) did not find any negative relationships between the Workload/Difficulty of a course and SET scores. The current study adopted the validity hypothesis interpreting the data, as research has additionally revealed that 'bias' factors had little and not consistent influence on SET scores (Addison et al. 2006; Centra 2003; Marsh and Roche 2000).

Marsh (1987) and Marsh and Roche (1997) described four background variables as the most influential: (a) Prior Subject Interest, (b) Expected Grade, (c) Workload/Difficulty, and (d) students selecting the course for General Interest. The design of the current study included the first three and the variable 'Professor Rank'. The 'Professor Rank' was selected because some earlier studies provided contradictory results about it (Feldman 1983; Marsh 2007a; Marsh and Hocevar 1991b; Spooren 2010). The exclusion of the 'General Interest' factor was decided because the majority of the courses evaluated were not a subject of free choice of the students, due to the structure of their program study.

Purpose of the study

The main goal of the current study was to provide a valid and reliable instrument for the evaluation of teaching effectiveness in the Greek higher education. Additional objectives were (a) to examine the dimensionality and the higher-order structure of the Greek version of SEEQ and (b) to investigate the effects of several background variables on SET scores provided by the Greek version of SEEQ.

Methods

Participants

A total of 1,264 students from 41 different courses of 14 departments participated in the current study by filling in the questionnaires (SEEQ) administered to them, between the tenth and the 13th week of the spring semester of 2013. The participants were students from the 14 out of 18 departments of Primary and Pre-primary Education of the Greek universities. The questionnaires of seven courses were excluded from further analysis due to low participation (under 13 students), as reliable results can be provided based on at least 10–15 student responses (Marsh 1982b). In Greece, students are not obliged to attend the courses. This was the reason that the selection of the courses relied mainly on compulsory courses of the curriculum where the attendance is higher than in optional courses, and more participation to the study had been expected. Indeed, the seven courses that were dropped from further analysis due to low participation were all optional courses. One additional reason for selecting departments only from Departments of Education was the differentiation detected in the SET scores among various disciplines (e.g. natural sciences, humanities, education courses) (Feldman 1978; Theall and Franklin 2001).

Instruments

The instrument used was the *Students' Evaluations of Educational Quality* (SEEQ) questionnaire (Marsh 1982b, 1987, 1991b). The SEEQ comprises 35 items for measuring nine dimensions (Learning/Value—five items, Instructor Enthusiasm—five items, Organization—four items, Group Interaction—four items, Individual Rapport—four items, Breadth of Coverage—four items, Examinations/Grading—three items, Assignments/Readings—two items, and Workload/Difficulty—four items).

Prior to the study, SEEQ was translated into the Greek language by the authors. Then, a bilingual academic with great experience in the field translated the instrument into English again. The bilingual academic and the authors compared the two versions of the instrument (the original and the back-translated version), and any discrepancies found were corrected in the Greek version of the instrument. Next, the translated SEEQ was administered to some students and colleagues as a pilot test of face validity. Some changes were introduced in order to improve the meaning of some items in Greek.

Procedure

The period for the administration of the questionnaires was prearranged and while students were filling in the questionnaires, the instructor was always outside the classroom. Afterwards, a student collected the questionnaires and put them in an envelope. The sealed envelope was then delivered by post to the authors. The students were informed that the

instructor would not receive feedback about the evaluation of his/her teaching before the end of the semester and that this feedback would be only summarized comments.

Models tested

A confirmatory factor analysis model (CFA) was conducted. Moreover, three higher-order CFA models were postulated and tested based on the study of Marsh (1991b), namely SEEQ Model 1 with one (general effectiveness), SEEQ Model 2 with two (skill and rapport), and SEEQ Model 3 with three (presenter, rapport, and regulator) higher-order factors. These higher-order CFA models are all nested in the initial factor model, having each a first-order factor load exclusively on a single higher-order factor.

Path analysis for the effect of background variables

The effect of a set of background variables on the SET scores of the Greek version of SEEQ was examined with path analysis. The set of the background variables that were included in the path model were Prior Student Interest in the Subject, Expected Grade, Professor Rank, and Workload/Difficulty. In the path model, the nine factors of the Greek version of SEEQ were included as latent variables.

Statistical analysis

CFA using Mplus v. 6.0 (Muthén and Muthén 1998–2010) was employed to confirm the factor structure and to examine the viability of three postulated higher-order models of the Greek version of the SEEQ. A path analysis was used to investigate the effects of several background variables on SET scores provided by the Greek version of SEEQ. In all models, the maximum likelihood estimator with robust standard errors was used, due to non-normality of several items.

To evaluate the factor structure of the Greek version of SEEQ, the following fit indices were used: (a) the root-mean-square error of approximation (RMSEA) which is relatively insensitive to sample size and has better index to test the model fit compared with χ^2 (Zhao and Gallant 2012), (b) the standardized root-mean-square residual (SRMR), and (c) the comparative fit index (CFI). The cut-off values for the goodness of fit for the above indices were based on Hu and Bentler (1999) suggestions. Although they acknowledged that it is difficult to define specific cut-off value for any index, they indicated that values for CFI around .95 for SRMR $\leq$ .08 and for RMSEA $\leq$ .06 seem to be appropriate for determining goodness of fit.

To compare the three higher-order models with the initial CFA model, besides the aforementioned indices (RMSEA, SRMR, and CFI), the Sattora–Bentler scaled Chi-square (Satorra and Bentler 2001) and the sample size adjusted BIC were used.

The effects of the background variables (Prior Student Interest in the Subject, Expected Grade, Professor Rank, and Workload/Difficulty) on the factors of the Greek version of the SEEQ were tested using path analysis.

Results

Confirmatory factor analysis of the Greek version of SEEQ

CFA results for the Greek version of SEEQ revealed acceptable values of the indices, supporting the proposed initial nine-factor structure of the original scale (Table 1).

Although the indices revealed reasonable fit to the data (initial model of the Greek SEEQ, Table 1), it was decided to try to further improve the fit by adding the terms that the modification indices suggested (final model of the Greek SEEQ, Table 1). The factor structure remained exactly the same in the initial and final model; the modification indices concerned only covariances between items of the scale. The model was further improved as it is revealed by the improved values of all fit indices presented in Table 1.

Higher-order structure of the Greek version of SEEQ

CFA results for the viability of three postulated models are presented in Table 2. The structure of the three higher-order models (Model 1, Model, 2, and Model 3) is based on the structure of the first-order model, without the additional parameters that the modifications indices proposed (i.e. initial model of the Greek SEEQ). This model sets the upper limit for the fit of the subsequent higher-order models (Marsh 1991b). The initial model of the Greek SEEQ (and not the final model of the Greek SEEQ) was chosen for comparison purposes, following the analysis of the Marsh study (1991b). Furthermore, a second reason was that the higher-order models would inherently include covariances between factors; thus, the covariances that have been added in the final model of the Greek SEEQ would be redundant. Each of the higher-order models was compared to the first-order base model using the Sattora–Bentler test (Table 2).

Model 1 has one higher-order factor, namely *general effectiveness* that consists of all nine first-order factors. Model 2 includes two higher-order factors, namely *rapport* (formed by Individual Rapport and Group Interaction), and *skill* (formed by the remaining seven first-order factors). Model 3 has three higher-order factors, namely *rapport* (formed by Individual Rapport and Group Interaction), *regulator* (formed by the Examinations/Grading, Assignments/Readings, and Workload/Difficulty), and *presenter* (formed by Learning/Value, Instructor Enthusiasm, Organization, and Breadth of Coverage).

All three higher-order models revealed acceptable values for the indices used compared to the initial model of the Greek SEEQ. Model 3 (three higher-order factors) appeared to be better than the remaining models positing one or two higher-order factors. Yet, the Sattora–Bentler scaled Chi-square, that was used to compare the initial CFA model with the three higher-order models, showed that in all three cases, the null hypothesis was rejected, supporting the better fit of the first-order CFA model over the competitive models. The Greek version of SEEQ higher-order structure differentiated slightly from the structure reported in Marsh's study (1991b), as a four higher-order factor structure was not supported. In that study (March 1991b), the three higher-order factor model was also a tenable model, yet slightly less sound than a four higher-order factor model.

Background variable effects on SET

Path analysis was adopted to investigate the role of background variables on SET scores. The set of background variables comprised the variables Prior Student Interest in the

Table 1 The fit of the Greek version of SEEQ

	χ^2 (*df*)	Sample size-adjusted BIC	CFI	RMSEA	SRMR
Initial model Greek SEEQ	1,934.382 (491)	133,948.583	.932	.048	.043
Final model Greek SEEQ	1,455.171 (486)	133,302.781	.954	.040	.039

Table 2 Model comparison between first-order CFA and SEEQ with one (Model 1), two (Model 2), and three (Model 3) higher-order factors

	χ^2 (*df*)	Scale corr. factor for MLR	Sample size-adjusted BIC	CFI	RMSEA	SRMR	Sattora–Bentler scaled χ^2 (*df*)
Initial model	1,934.382 (491)	1.365	133,948.583	.932	.048	.043	
Model 1	2,339.750 (518)	1.372	134,410.571	.914	.053	.056	379.982 (27)
Model 2	2,285.658 (517)	1.371	134,339.184	.916	.052	.055	332.280 (26)
Model 3	2,189.774 (515)	1.370	134,212.757	.921	.051	.054	244.217 (24)

Subject, Expected Grade, Professor Rank, and Workload/Difficulty, whereas the nine factors of the Greek version of SEEQ served as the latent factors of the path model. The results supported the fit of the model ($\chi^2 = 1{,}663.232$, $df = 567$, CFI = .950, RMSEA = .04, and SRMR = .039).

Expected Grade affected positively all the nine factors of the Greek version of SEEQ. Professor Rank affected negatively the factors Instructor Enthusiasm, Group Interaction, and Individual Rapport, positively the factors Learning/Value, Organization, and Assignments/Readings, whereas no significant effects were detected on the factors Breadth of Coverage, Examinations/Grading, and Workload/Difficulty. Prior Student Interest in the Subject affected positively all the factors of the Greek version of SEEQ, whereas no significant effect was detected on the Workload/Difficulty. Finally, Workload/Difficulty affected the factor Learning/Value negatively, the factor Examinations/Grading positively, whereas no significant effect was detected for the other factors of the Greek version of SEEQ. The results of the path analysis are presented in Fig. 1.

Discussion

Psychometric properties of the Greek version of SEEQ

The basic aim of this study was to provide the Greek higher education setting with a valid and reliable instrument for the evaluation of teaching effectiveness. To this respect, the factorial validity of the Greek version of SEEQ was supported. The 35 items of the questionnaire comprised a nine-factor model exactly the same as the original SEEQ (Marsh 1982b). The results added additional support to the strong psychometric properties of SEEQ across different cultures and environments (Balam and Shannon 2010; Coffey and Gibbs 2001; Marsh 1986; Marsh et al. 1997; Watkins and Thomas 1991). The findings are in accordance with the widely supported notion that SET instruments should capture multiple dimensions of teaching effectiveness (Spooren et al. 2013).

The provision of a valid teaching effectiveness instrument will help Greek researchers to overcome the limited value and usefulness of several 'home-made' instruments used currently in Greece. 'Home-made' instruments are usually developed without relying on a theory of teaching and learning and do not use a rigorous approach that guarantees validity (Marsh 2001). Committees or university faculty staff members, who construct such instruments, do not usually evaluate them in terms of psychometric properties and revise them accordingly. Yet, SET instruments measuring distinct elements of teaching

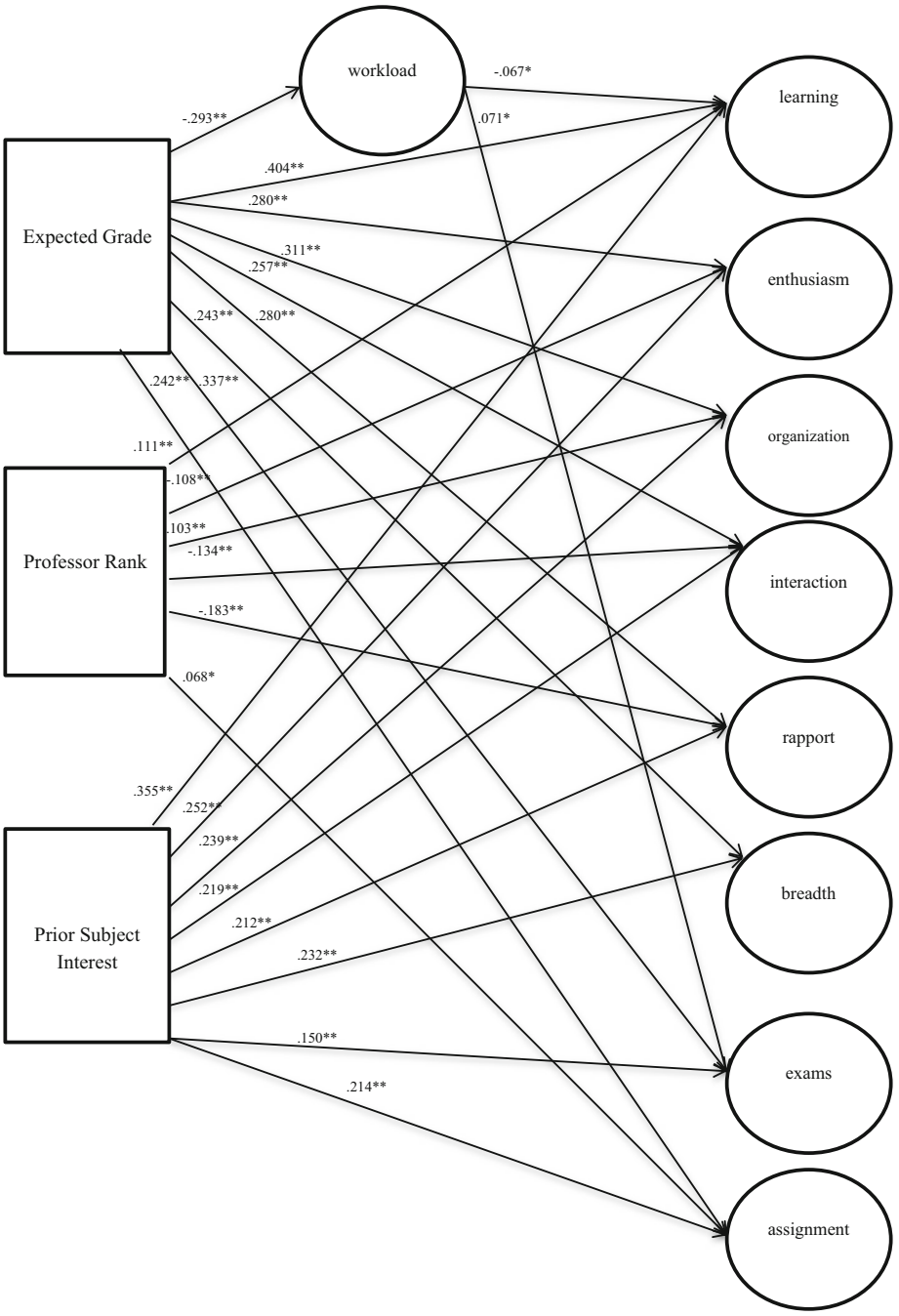

Fig. 1 Path diagram: effects of the background factors on the scores of the Greek version of SEEQ. *significance level at .05, **significance level at .01

effectiveness have to be controlled for both content and construct validity (Marsh 2001). To this respect, the Greek version of SEEQ is an easy to use measure with sound psychometric properties and can serve as a basis for comparable findings in the Greek higher education settings.

Multidimensionality and higher-order structure

While other methods, such as peer evaluation, committee of experts, or student learning outcomes, are available to evaluate teaching effectiveness, SET still dominate the higher education settings (Balam and Shannon 2010). In his seminal study, Marsh (1982b) argued that students' evaluations of teaching effectiveness are best understood by multiple dimensions, as teaching shall be regarded a multidimensional construct. The results of this study confirmed the multidimensionality of the teaching effectiveness construct evaluated by the Greek version of SEEQ.

The wide variety of dimensions captured in different SET instruments initiated the discussion about the existence of a single higher-order factor representing the teaching effectiveness contrast (Spooren et al. 2013). Additional pressure towards this direction was added by the use of SET scores for summative and administration purposes, where a single score capturing a general instructional skill would be very meaningful (Apodaca and Grad 2005). Abrami (1985) argued for a single higher-order factor model, whereas other research studies supported a multiple higher-order factors structure of teaching effectiveness (Apodaca and Grad 2005; Burdsal and Harrison 2008; Cheung 2000; Harrison et al. 2004; Marsh 1991b; Mortelmans and Spooren 2009). The current findings provide further evidence against the single higher-order structure (Model 1) of the teaching effectiveness construct, revealing that a three higher-order factor model (Model 3) to be most tenable, as it was shown by the Sattora–Bentler test.

Additional research is expected to shed light in the number of the higher-order factors of the Greek version of SEEQ. The results of the current study supported the notion that even if a single score for summative purposes is more useful, SET scores are still multidimensional. A compromise is yet possible if weighted scores are going to be used (Marsh 1991b).

The effects of background variables

As it was mentioned before, the current study adopted a construct validity approach. The leniency hypothesis was not supported by the results, as the Workload/Difficulty dimension has no significantly effect on six out of eight dimensions of the Greek version of SEEQ. Workload/Difficulty significantly affected only the factors Learning/Value ($-.067$, $p < .05$) and Examinations/Grading ($.071$, $p < .05$), but these low values could not support the leniency hypothesis. The negative correlation between Workload/Difficulty and Learning/Value might be attributed to the fact that students with excessive workload are unlikely to absorb the provided material and as a result, learning 'suffers' (Marsh and Roche 2000). This positive effect of the Workload/Difficulty on Examinations/Grading could be attributed to the fact that students' evaluation in 'difficult' courses might be better designed and implemented by the instructor. Yet, due to the low values, it is premature to draw any firm conclusions, and further investigation of the Workload/Difficulty role shall be an objective of future studies.

The Expected Grade's significant effects on the variables of the Greek version of SEEQ also supported the validity hypothesis. The highest effect was on Learning/Value variable ($.404$, $p < .01$) indicating that better expected grades might reflect better learning by

students. It has to be pointed here that these two concepts are not equated but that Expected Grade reflects only a measure of learning (Marsh and Roche 1997). Additional support was revealed by the negative correlation between Expected Grade and Workload/Difficulty ($-.293$, $p < .01$). This is an a priori reasonable prediction, as students' expectations for grades are unlikely to be high due to the difficulty and heavy load of a specific course.

In our study, Prior Subject Interest mostly influenced Learning/Value as in Marsh and Roche (1997) study. Moreover, its positive effect on eight dimensions of the Greek version of SEEQ supported the validity hypothesis, as students with higher prior interest in the course shall be expected to provide more positive SET ratings (Marsh and Roche 1997). The only factor that was not significantly correlated with Prior Subject Interest was the Workload/Difficulty factor. This is also an a priori reasonable prediction, as the interest in an academic subject and the difficulty of a course are unlikely to be correlated.

The fourth background variable tested was Professor Rank. It can be assumed that younger instructors (lower rank) would be more enthusiastic, provide better interactions, and create a more positive climate than the older ones (higher rank) (Feldman 1983). Indeed, the current study's findings revealed negative correlations between Professor Rank and Enthusiasm, Group Interaction, and Individual Rapport. On the other hand, the positive correlations with Learning/Value, Organization, and Assignments/Readings can be attributed to the accumulated experience of higher-ranked Professors. These results are in contradiction with Marsh and Hocevar (1991b) and Marsh (2007a) studies, where added experience did not influence teachers' effectiveness. To this respect, further research efforts are needed to explore the relation between the staff's experience and SET effectiveness in the Greek educational settings.

In conclusion, the current study has theoretical and practical importance. This study provided solid evidence for the applicability of SEEQ in the Greek higher education, by confirming the factor structure of the instrument. The intention of this study was to provide the Greek higher education setting with a valid instrument that can be used in a variety of environments for internal evaluation purposes. Moreover, the results reassured the sound psychometric properties of SEEQ indicating the value and usefulness of the instrument. This explains why a large part of the discussion was focused on psychometric issues. Future studies, with a more representative and diverse sample, will allow a more analytic and in-depth elaboration regarding the actual results of the instrument.

Limitations of the study

The structure of the program of study in the Greek universities and the limited sources in terms of financial support were responsible for some limitations of the current study.

a. The role of the important background factor of students selecting a course for "General Interest" (Marsh 1987, Marsh and Roche 1997) could not be incorporated in the current study because of the compulsory nature of the most courses selected in the current study. Additionally, the sample of the current study relied only on courses at School of Education departments.
b. Several studies revealed the importance of multiple sources of information to better capture the SET construct. Yet, the findings of the current study relied only on the use of students' evaluations of teaching effectiveness, as no other measure evaluating teaching effectiveness was used (e.g. teacher evaluations, peer evaluation, panel of experts).

c. The underlying structure of the Greek version of SEEQ was based on the back translation and the CFA adopted in the current study, evaluating only the face and structural validity.

Conclusions

Evaluating teaching effectiveness in higher education is not an easy task due to the complexity of the construct. Even if the results of the current study were promising, additional research will help to better understand the SET construct in the Greek higher educational settings. Future efforts must select a wider variety of departments towards a representative sample, investigate further the role of more background factors of the teaching effectiveness construct, and elaborate more by a deeper analysis of the results. Additional resources evaluating teaching effectiveness will shed more light on the SET construct by providing comparable data. Finally, future research studies must further investigate others aspects of validity such as 'evaluating test validity is not a static, one-time event; it is a continuous process' (Mertens 2010, p. 384).

References

Abrami, P. C. (1985). Dimensions of effective college instruction. *Review of Higher Education, 8*(3), 211–228.

Abrami, P. C., & d'Apollonia, S. (1991). Multidimensional students' evaluation of teaching effectiveness-generalizabity of 'N = 1' research. Comment on Marsh (1991). *Journal of Educational Psychology, 83*(4), 411–415.

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

Abrami, P. C., & Mizener, D. A. (1983). Does the attitude similarity of college professors and their students produce 'bias' in course evaluations? *American Educational Research Journal, 20*(1), 123–136.

Addison, W. E., Best, J., & Warrington, J. D. (2006). Students' perceptions of course difficulty and their ratings of the instructor. *College Student Journal, 40*(2), 409–416.

Aleamoni, L. M. (1987). Student rating myths versus research facts. *Journal of Personnel Evaluation in Education, 1*(1), 111–119.

Aleamoni, L. M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education, 13*(2), 153–166.

Alsmadi, A. (2005). Assessing the quality of students' ratings of faculty members at Mu'tah University. *Social Behaviour and Personality, 33*(2), 183–188.

Apodaca, P., & Grad, H. (2005). The dimensionality of student ratings of teaching: Integration of uni- and multidimensional models. *Studies in Higher Education, 30*(6), 723–748.

Balam, E. M., & Shannon, D. M. (2010). Student ratings of college teaching: a comparison of faculty and their students. *Assessment & Evaluation in Higher Education, 35*(2), 209–221.

Boysen, G. A., Kelly, T. J., Raeslyand, H. N., & Casner, R. W. (2014). The (mis)interpretation of teaching evaluations by college faculty and administrators. *Assessment & Evaluation in Higher Education, 39*(6), 641–656.

Burdsal, C. A., & Harrison, P. D. (2008). Further evidence supporting the validity of both a multidimensional profile and an overall evaluation of teaching effectiveness. *Assessment & Evaluation in Higher Education, 33*(5), 567–576.

Cashin, W. E. (1989). *Student ratings of teaching: Recommendations for use*. (IDEA Paper no. 22). Manhattan, KS: Kansas State University, Center for Faculty Evaluation and Development.

Cashin, W. E. (1995). *Student ratings of teaching: The research revisited*. (IDEA Paper no. 32). Manhattan, KS: Kansas State University, Center for Faculty Evaluation and Development.

Centra, J. A. (2003). Will teachers receive higher student evaluations by giving higher grades and less course work? *Research in Higher Education, 44*(5), 495–518.

Centra, J. A., & Gaubatz, N. B. (2000). Is there gender bias in student evaluations of teaching? *The Journal of Higher Education, 71*(1), 17–33.

Cheung, D. (2000). Evidence of a single second-order factor in student ratings of teaching. *Structural Equation Modeling, 7*(3), 442–460.

Chism, N. V. N. (1999). *Peer review of teaching: A sourcebook*. Bolton, MA: Anker Pub. Co.

Coffey, M., & Gibbs, G. (2001). The Evaluation of the Student Evaluation of Educational Quality Questionnaire (SEEQ) in UK Higher Education. *Assessment & Evaluation in Higher Education, 26*(1), 89–93.

d'Apollonia, S., & Abrami, P. C. (1997). Navigating student ratings of instruction. *American Psychologist, 52*(11), 1198–1208.

El Hassan, K. (2009). Investigating substantive and consequential validity of student ratings of instruction. *Higher Education Research and Development, 28*(3), 319–333.

Feely, T. H. (2002). Evidence of halo effects in student evaluations of communication instruction. *Communication Education, 51*(3), 225–236.

Feldman, K. A. (1976). The superior college teacher from the student's view. *Research in Higher Education, 5*(3), 243–288.

Feldman, K. A. (1978). Course characteristics and college students' ratings of their teachers and courses: What we know and what we don't. *Research in Higher Education, 9*(3), 199–242.

Feldman, K. A. (1983). The seniority and instructional experience of college teachers as related to the evaluations they receive from their students. *Research in Higher Education, 18*(1), 3–124.

Fincher, C. (1985). Learning theory and research. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research* (pp. 63–96). New York, NY: Agathon Press.

Germain, M. L., & Scandura, T. A. (2005). Grade inflation and student individual differences as systematic bias in faculty evaluations. *Journal of Instructional Psychology, 32*(1), 58–67.

Harrison, P., Douglas, D., & Burdsal, C. (2004). The relative merits of different types of overall evaluations of teaching effectiveness. *Research in Higher Education, 45*(3), 311–323.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55.

Jones, J., Gaffney-Rhys, R., & Jones, E. (2014). Handle with care! An exploration of the potential risks associated with the publication and summative usage of student evaluation of teaching (SET) results. *Journal of Further and Higher Education, 38*(1), 37–56.

Langbein, L. (2008). Management by results: Student evaluation of faculty teaching and the mis-measurement of performance. *Economics of Education Review, 27*(4), 417–428.

Marsh, H. W. (1982a). The use of path analysis to estimate teacher and course effects in student ratings of instructional effectiveness. *Applied Psychological Measurement, 6*(1), 47–59.

Marsh, H. W. (1982b). SEEQ: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology, 52*(1), 77–95.

Marsh, H. W. (1986). Applicability paradigm: Students' evaluations of teaching effectiveness in different countries. *Journal of Educational Psychology, 78*(6), 465–473.

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research, 11*(3), 253–388.

Marsh, H. W. (1991a). A multidimensional perspective on students' evaluations of teaching effectiveness: Reply to Abrami & d'Apolonia (1991). *Journal of Educational Psychology*, *83*(3), 416–421.

Marsh, H. W. (1991b). Multidimensional students' evaluations of teaching effectiveness: A test of alternative higher-order structures. *Journal of Educational Psychology, 83*(2), 285–296.

Marsh, H. W. (1994). Weighting for the right criteria to validate student evaluations of teaching in the IDEA system. *Journal of Educational Psychology, 86*(4), 631–648.

Marsh, H. W. (1995). Still weighting for the right criteria to validate student evaluations of teaching in the IDEA system. *Journal of Educational Psychology, 87*(4), 666–679.

Marsh, H. W. (2001). *Students' evaluations of university teaching*. Paper presented as part of an invited lecture and workshop presentation in University of Braga, Portugal on June 2001 (Retrieved after personal communication with Prof. Marsh on March 2012).

Marsh, H. W. (2007a). Do university teachers become more effective with experience? A multilevel growth model of students' evaluations of teaching over 13 years. *Journal of Educational Psychology, 99*(4), 775–790.

Marsh, H. W. (2007b). Students' evaluations of university teaching: A multidimensional perspective. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319–384). New York, NY: Springer.

Marsh, H. W., & Dunkin, M. J. (1992). Students' evaluations of university teaching: A multidimensional perspective. In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 241–320). New York, NY: Agathon.

Marsh, H. W., Hagu, K. T., Chung, C. M., & Siu, T. L. P. (1997). Students' evaluations of university teaching: Chinese version of the students' evaluations of educational quality (SEEQ) instrument. *Journal of Educational Psychology, 89*(3), 568–572.

Marsh, H. W., & Hocevar, D. (1991a). The multidimensionality of students' evaluations of teaching effectiveness: The generality of factor structures across academic discipline, instructor level, and course level. *Teaching and Teacher Education, 7*(1), 9–18.

Marsh, H. W., & Hocevar, D. (1991b). Students' evaluations of teaching effectiveness: The stability of mean ratings of the same teachers over a 13-year period. *Teaching and Teacher Education, 7*(4), 303–314.

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective. *American Psychologist, 52*(11), 1187–1197.

Marsh, H. W., & Roche, L. A. (2000). Effects of grading leniency and low workload on students' evaluations of teaching: Popular myth, bias, validity, on innocent bystanders? *Journal of Educational Psychology, 92*(1), 202–228.

McCann, S., & Gardner, C. (2014). Student personality differences are related to their responses on instructor evaluation forms. *Assessment & Evaluation in Higher Education, 39*(4), 412–426.

McKeachie, W. J. (1997). Student ratings: The validity of use. *American Psychologist, 52*(11), 1218–1225.

McPherson, M. A., & Jewell, R. T. (2007). Leveling the playing field: Should student evaluation scores be adjusted? *Social Science Quarterly, 88*(3), 868–881.

Mertens, D. M. (2010). *Research and evaluation in education and psychology. Integrating diversity with quantitative, qualitative, and mixed methods* (3rd ed.). Thousand Oaks, CA: Sage.

Mortelmans, D., & Spooren, P. (2009). A revalidation of the SET37-questionnaire for student evaluations of teaching. *Educational Studies, 35*(5), 547–552.

Muthén, L. K., & Muthén, B. O. (1998–2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.

Safer, A. M., Farmer, L. S. J., Segalla, A., & Elhoubi, A. F. (2005). Does the distance from the teacher influence student evaluations? *Educational Research Quarterly, 28*(3), 28–35.

Satorra, A., & Bentler, P. M. (2001). A scaled difference Chi-square test statistic for moment structure analysis. *Psychometrika, 66*(4), 507–514.

Spooren, P. (2010). On the credibility of the judge. A cross-classified multilevel analysis on students' evaluation of teaching. *Studies in Educational Evaluation, 36*(4), 121–131.

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research, 83*(4), 598–642.

Theall, M., & Franklin, J. (2001). Looking for bias in all the wrong places: A search for truth or a witch hunt in student ratings of instruction? *New Directions for Institutional Research, 2001*(109), 45–56.

Watkins, D., & Thomas, B. (1991). Assessing teaching effectiveness: An Indian perspective. *Assessment & Evaluation in Higher Education, 16*(3), 185–198.

Zabaleta, F. (2007). The use and misuse of student evaluation of teaching. *Teaching in Higher Education, 12*(1), 55–76.

Zhao, J., & Gallant, D. J. (2012). Student evaluation of instruction in higher education: Exploring issues of validity and reliability. *Assessment & Evaluation in Higher Education, 37*(2), 227–235.