Journal of Educational and Psychological Studies - Sultan Qaboos University (Pages 656-667) Vol.14 Issue 4, 2020
DOI: http://dx.doi.org/10.24200/jeps.vo14iss4pp656-667

Psychometric Properties of an Instrument Developed to Assess Students' Evaluation of Teaching in Higher Education

Mutasem M. Akour* & Bashar K. Hammad
The Hashemite University, Jordan German Jordanian University, Jordan

Received: 1/7/2020 Accepted: 6/9/2020

Abstract: Student evaluation of teaching is a global predominant practice in higher education institutions. Therefore, a major university in Jordan developed a questionnaire for students' use in evaluating their instructors' teaching effectiveness. Since student evaluation of teaching is an important process, the present study tried to examine the psychometric properties of the instrument. Item-total correlations showed acceptable internal consistency. In addition, a two-factor structure of the scale (teaching effectiveness and course attributes) was supported by exploratory factor analysis and confirmatory factor analysis from two independent samples. Convergent validity was supported by a moderate correlation coefficient between course averages of students' ratings on the first factor and course averages of students' final grades in each course. Finally, students' responses on the factor that captures teaching effectiveness were found to have very high internal consistency (Cronbach's alpha of 0.96). However, this instrument lacks evidences of content validity and convergent validity. Therefore, it is important to be cautious in evaluating faculty members and making promotion decisions that is based solely on the scores obtained using this instrument.

Keywords: Students' ratings, validity, reliability, teaching effectiveness.

الخصائص السيكومترية لمقياس تقييم طلبة الجامعة لفاعلية العملية التدريسية

معتصم عكور* وبشار حماد
الجامعة الهاشمية، الأردن الجامعة الالمانية الاردنية، الاردن

مستخلص: تعتبر عملية تقييم الطلبة لفاعلية العملية التدريسية من الممارسات الشائعة في مختلف الجامعات حول العالم، لذلك قامت إحدى الجامعات الاردنية بتطوير مقياس يستخدمه الطلبة في تقييم فاعلية العملية التدريسية لاعضاء هيئة التدريس فيها. ونظراً لأهمية عملية تقييم الطلبة لفاعلية العملية التدريسية، هدفت الدراسة الحالية الى التحقق من الخصائص السيكومترية لهذا المقياس. أظهرت النتائج الخاصة بصدق البنية الداخلية للمقياس أن قيم معامل الارتباط المصحح بين الفقرة والاداة كانت مقبولة ، في حين أشارت نتائج كلا من التحليل العاملي الاستكشافي والتوكيدي أن الفقرات تشبعت على عاملين: العامل الاول ويمثل فاعلية العملية التدريسية ، والعامل الثاني ويمثل خصائص المساق. وفيما يتعلق بالصدق التقاربي للمقياس، فقد تم التحقق منه من خلال إيجاد معامل الارتباط بين متوسطات تقديرات الطلبة على كل مساق مع متوسطات علامات الطلبة في تلك المساقات. وأخيراً، فقد تم تقديرالثبات من خلال طريقة الاتساق الداخلي باستخدام معامل الفا لكرونباخ حيث كانت قيمته مرتفعة (٠.٩٦). نظرا لعدم وجود أي دلائل على صدق المحتوى وأي دلائل إضافية على الصدق التقاربي لهذه الاداة ، يجب أن يكون هناك حذر في استخدام هذه الاداة في تقييم أعضاء هيئة التدريس وفي إتخاذ أي قرارات تخص ترقيتهم بالاعتماد فقط على نتائجهم على هذه الاداة.

الكلمات المفتاحية: تقديرات الطلبة لفاعلية العملية التدريسية، الصدق، الثبات، طلبة اللجامعة.

*mutasem@hu.edu.jo

Even though student evaluation of teaching (SET) has been around since the mid-1920s, its use for both formative (e.g., as feedback for the improvement of teaching) and summative purposes (e.g., mapping teaching competence for administrative decision-making) started during the 1970s (Morley, 2014). Nowadays, SET is used in almost every institution of higher education throughout the world for improving teaching quality, making tenure/promotion decisions, providing information to students for the selection of courses and teachers, and providing evidence for institutional accountability (Spooren, Brockx, & Mortelmans, 2013; Zhao & Gallant, 2012).

Instructors are convinced of the merit of SET as a tool for feedback on their teaching, since students can provide an insight into the strengths and weaknesses regarding their instructors' teaching practices. Students can provide information about accomplishment of major educational objectives, instruction materials and instructional methods, kind of communication between students and the instructor, and rapport with the instructor (Balam & Shannon, 2009; Griffin, 2001) . Therefore, many instructors welcome the results of SET in order to improve their subsequent instruction.

On the other hand, some extraneous factors could affect students evaluations of teaching effectiveness that are outside of the instructor's control, such as: student motivation, student grade point avarage, class size, subject matter, and gender (Willits & Brennan, 2017). Moreover, many faculty members question the reliability and validity of SET results since these results can have serious effects on their professional career (Kogan, Schoenfeld-Tacher, & Hellyer, 2010; Ory, 2001). That is, faculty members have concerns about the extent to which students are capable of providing appropriate instructor evaluations due to the differences between the ways in which students and instructors perceive effective teaching (Spooren et al., 2013). Many instruments were not tested with regard to their psychometric properties, i.e., reliability and validity (Richardson, 2005).

Reliability quantifies the precision of scores obtained from a given measure. It is merely concerned with how scores resulting from an instrument would be expected to vary across replications of that instrument. Under classical test theory, two common methods for estimating the reliability coefficient are test-retest and internal consistency methods (Haertel, 2006). On the other hand, validity of a SET instrument is the extent to which scores generated by an instrument measure what is intended to measure, i.e., teaching effectiveness. In validation studies, researchers seek to provide evidences from the following three main types of evidences. First, content-related evidences, which concerns the extent to which the items of an instrument are appropriate representations of the content being measured. Second, internal-structure evidences, which concerns the factor structure of the instrument and the internal consistency within each factor, Third, association with other variables which concerns the extent to which scores are related to (convergent validity) or not related to (divergent validity) other variables (Kane, 2006; Onwuegbuzie, Daniel, & Collins, 2007).

Previous research on SET has focused on two general areas. The first one concerned the biasing of student ratings by different unrelated factors to teaching quality. The second one is examining the quality of SET instruments through collecting evidences of validity and reliability (Willits & Brennan, 2017).. In the current study we are limited to the second area.

Penny (2003) claimed that many of the SET instruments lack evidences of construct validity, and therefore it is legitimately to question any inferences and decisions drawn from them. Therefore, numerous studies in the literature were conducted to describe the development and validation of various SET instruments developed at different universities. For example, Nemec, Baker, Zhang, and Dintzner (2018) developed an evaluation tool to be used by a college of pharmacy at the Western New England University in

the United States. A total of 199 items were compiled from a review of the related literature and grouped into six subscales that was intended to evaluate the instructor and another six subscales that was intended to evaluate the course. Findings of this study showed that all subscales for the evaluation of instructor had high internal consistency where Cronbach's alpha was above 0.9, whereas it was above 0.8 for the course subscales. Moreover, Confirmatory factor analysis revealed a moderate model fit with factor loadings for all items above 0.6.

In another study that aimed at validating a local 17-item SET scale used at a large Italian university, Bassi, Clerci, and Aquario (2017) used data from 54,777 questionnaires in the academic year 2012-2013, and found out that item-total correlations were all above 0.60. Moreover, exploratory factor analysis resulted in a four-factor solution, with all item related to teaching effectiveness loaded on the same factor. Cronbach's alpha for the 17-item scale was 0.97, while it was 0.90 for the teaching effectiveness subscale.

Moreover, Oon, Spencer, and Kam (2016) investigated psychometric quality of a student evaluation of teaching survey developed at the University of Macau. The survey consisted of five items such that each item was designed to measure one dimension related to student learning and teaching. Confirmatory factor analysis confirmed the unidimensionality of the scale. Zhao and Gallant (2012) examined the validity and reliability of a 10-item instrument used at large mid-western university in the United States to measure students' evaluation of instruction using 73,500 students' responses. The findings of the study showed high internal consistency with a value of Cronbach's alpha of 0.95. In addition, a confirmatory factor analysis provided that a unidimensional model had an acceptable fit to the data.

In addition, Catano and Harvey (2011) developed and validated a SET instrument to be used in a Canadian University that consisted of nine teaching effectiveness competences. Cronbach's alpha was 0.94, and exploratory factor analysis proved that the measure is unidimensional. Moreover, a 0.30 correlation coefficient was found between scores on the SET measure and students' GPAs. Finally, Chen and Watkins (2010) developed a SET instrument at one teaching and research university in China. They used data from 7560 students to validate this instrument. Confirmatory factor analysis revealed a one-factoe structre, and Cronbach's alpha was 0.98.

It is obvious that some studies supported the multi-factor structure of the home-grown SET scales (Nemec et al., 2018), whereas many studies supported the one-factor structure (Bassi et al., 2017; Catano & Harvey, 2011; Chen & Watkins 2010 ; Oon et al., 2016; Zhao & Gallant, 2012). On the other hand, almost all studies yielded high (above 0.90) internal consistency estimates of reliability. In addition, most validation studies were conducted in the western countries, and no study was found in the Arabic context. Therefore, it is of importance to enrich the literature with studies that show the experiences of different universities in the Arab world concerning student evaluation of teaching effectiveness.

Statement of the problem

According to the increasing role of students in assessing instructors and instructional quality in most higher education institutions, the Hashemite University (HU) in Jordan incorporated the use of students' evaluation of teaching effectiveness in its education policy. The HU is one of the major state universities in Jordan that was established in 1995. It comprises 16 faculties and more than 50 departments (for more information, visit www.hu.edu.jo). The process of evaluating teaching effectiveness at the HU started in the first semester of the academic year 2004/2005, when it was carried out manually.

The existing SET instruments in the literature vary greatly in both content and construction, due to the characteristics and desires of an institution (Spooren et al., 2013). Given this, Spooren et al. recommended that institutions should be able to select the aspects that are most important, according to their educational vision and policy, thereby developing SET instru-

ments that are consistent with their own preferences. Therefore, the HU developed an instrument to be used by its student in evaluating the teaching effectiveness of their instructors. The purpose of the current study was to validate this instrument by exploring its factor structure, and by investigating reliability of ratings.

Significance of the study

This study was motivated by the fact that HU relies on the results of SET in making decisions related to faculty retention, promotion and merit pay. Therefore, it is important to investigate the psychometric properties of such an existing instrument that has been used in HU for several years. It is hoped that the results of the currents study would provide more evidences about the credibility of using this instrument in the improvement of teaching quality, and for administrative decision-making purposes such as awarding tenure and promotion.

This study was also motivated by the lack of validated SET questionnaires in the Arabic context. Studies about students' evaluation of teaching is not scarce. However, the majority of previous studies (e.g., Aleamoni, 1978; Bassi et al., 2017; Catano & Harvey, 2011; Lemos, Queirós, Teixeira, & Menezes, 2011; Marsh , 1982; Nemec et al., 2018; Wilson, Lizzio, & Ramsden, 1997) were conducted in the western countries.

It is hoped that validating this instrument would be useful in providing other Arabic-speaking institutions with a psychometrically sound instrument that can be used in evaluating teaching effectiveness in their institutions, and in encouraging them in investigating the psychometric properties of the instruments they already use.

Method

Data

The current study was based on data collected by the Information, Communication & E-learning Technology Center (ICET) at the HU in the first and second semesters of the academic year 2012/2013. Data for the first semester consisted of 6,105 completed responses on the evaluation instrument in more than 400 different courses. Data from 904 participants were eliminated because of extreme responding (i.e., choosing only 1s or 5s), resulting in a final sample size of 5,201. Approximately 62% of participants in the final sample were females. About half (52%) of respondents had a grade-point-average of 2.75 and above, which translates to a letter grade of B- and above. Furthermore, the majority (60%) of records represented responses for freshmen, 6% for sophomores, 14% for juniors, and 20% for seniors.

Moreover, data for the second semester of the academic year 2012/2013 consisted of 5,956 completed responses on the evaluation instrument in more than 400 different courses. Data from 420 participants were eliminated because of extreme responding (i.e., choosing only 1s or 5s), resulting in a final sample size of 5,536. Approximately 61% of participants in the final sample were females. About half (53%) of respondents had a grade-point-average of 2.75 and above, which translates to a letter grade of B- and above. Furthermore, the majority (58%) of records represented responses for freshmen, 6% for sophomores, 15% for juniors, and 21% for seniors.

Instrument

In 2010, a committee was formed to study comments from faculty members and students on the evaluation process. As a result, a new evaluation questionnaire was developed including questions about different elements of the teaching process. The preference at HU was to develop a parsimonious instrument that captures the dimensions of effective teaching (e.g., effective communication of the material to students, and positive interactions with the students) which are most important according to the educational vision and policy of the HU.

The committee developed an initial item pool of 33 items using the literature and other questionnaires developed by different well-known universities in the United States (e.g., the Arizona Course/Instructor Evaluation Questionnaire (Aleamoni, 1978), the Students' Evaluation of Education Quality (Marsh, 1982), the Course Ex-

perience Questionnaire (Ramsden, 1991; Wilson et al., 1997), the Students' Evaluation of Teaching Effectiveness Rating Scale (Toland & De Ayala, 2005), and the Pedagogical Questionnaire of the University of Porto (Lemos et al., 2011).

The interest was on developing an instrument that applies to all programs, focuses on teaching quality, and that is short and less time-consuming. After several meetings and extensive brainstorming, the committee selected 18 items from the initial item pool to represent the evaluation instrument. These items were sought to evaluate two constructs, the teacher, and the course. The first 14 items asked about the instructor: use of appropriate teaching methods and teaching aids (items 7, and 13), subject knowledge (items 4 and 6), office hours (item 8), assessment of students (items 11 and 12), interaction with students (items 1, and 3), fulfillment of course objectives (items 5, 9, and 10), classroom management (item 2), and overall judgment (item 14). The last four questions asked about course materials (item 15 and 17), atmosphere of the class (item 18), and recommendation for forthcoming students (item 16). All items were rated on a five-point Likert-type rating scale, ranging from 1 (strongly disagree) to 5 (strongly agree).

The final form of the questionnaire consisted of two dimensions: teaching effectiveness which was represented by 14 items, and course attributes which was represented by 4 items. This final form was converted to a web-based form where all students have the access through their portals to complete it securely and confidentially. This form was administered to students during the last two weeks of class in the semester. Students' responses to the first 14 items were averaged to produce a mean score, which is used as an index of teaching. effectiveness. Demographic information such as students' gender, grade point average, year of study, and other information were gathered electronically through the ICET at the HU.

Data Analysis

Descriptive statistics, mean and standard deviation, were computed for the scores on each item of the instrument that was designed to measure teaching effectiveness. In addition, corrected-item total correlations were computed as an indicator of internal consistency of the instrument. Values less than 0.3 indicates that an item does not measure teaching effectiveness and should be eliminated (Nunnally & Bernstein, 1994).

To explore factor structure of the instrument, principal component analysis with direct oblimin rotation was conducted on the 18-item scale using SPSS 25. The oblique method of rotation is recommended when items are assumed to be correlated with each other (Field, 2009). Given the findings of previous research (Benton & Cashin, 2014; Martínez-Gómez, Sierra, Jabaloyes, & Zarzo, 2010; Ramsden, 1991) of positive and moderate intercorrelations between different dimensions of the developed scales, an oblique rotation was assumed the most appropriate.

For cross validation of the results, Confirmatory Factor Analysis (CFA) was conducted on the data for the second semester using LISREL 8.80 program (Jöreskog & Sörbom, 2006). The following indices were used to assess model's fit to the data: chi-square (χ^2), root mean square error of approximation (RMSEA), standardized root mean square residual (SRMR), and goodness-of-fit index (GFI). Usually a non-significant value of χ^2 indicates a model fit to the data. However, χ^2 is sensitive to sample size, and therefore is not as reliable as other indices of model fit (Schumacker & Lomax, 2010). The model will be considered adequate if the index GFI show values above 0.90, SRMR is below 0.05, and if the RMSEA is below .10 (Kelloway, 1998).

As a measure of convergent validity of the SET Instrument, authors used to correlate individual student ratings with the students' individual final course grades. However, correlating individual SET scores with individual final course grades is not appropriate as measure of validity of the instrument. To determine whether better evaluation results go along with better student learning outcomes, one needs to correlate course averages of SETs and course averages of examination

scores. A significant positive correlation between SETs and examination scores on the course level suggests that effective teaching affects both, student ratings and examination results. This would support the validity of SETs as measures of teaching effectiveness. When analysed on the individual student level, however, differences between SETs and test scores would need to be attributed to differences between individual students, e.g., individual differences regarding motivation or ability (Stehle, Spinath, & Kadmon, 2012). Finally, Cronbach's alpha was computed to estimate the reliability of the ratings in the instrument.

Results

Initial Item Analysis

The 18 items in the instrument, along with the mean, standard deviation, and corrected item-total correlation for each item are presented in Table 1. Corrected item-total correlations for all items fall above 0.30, and ranged from 0.38 for item 16 "I recommend my colleagues to study this course" to 0.85 for item 12 "Instructor's homework and exams conform to material covered in the course." Therefore, none of the items was eliminated in this initial item analysis.

As reflected in Table 1, the mean ratings ranged from 3.49 (SD= 1.39) to 4.32 (SD= 1.05) on a five-point scale. For the first 14 items that relate to the instructor, the mean ratings ranged from 3.82 (SD= 1.44) to 4.32 (SD= 1.05). However, for the last four items the mean ratings ranged from 3.49 (SD= 1.39) to 3.90 (SD= 1.27). All items tended to have comparable high mean students' ratings as shown in Figure 1.

Table 1
Mean, standard deviation, and corrected item-total correlation for the SET items (N= 5,201)

Item	*M*	SD	Corrected item-total correlations
1- Mutual respect predominates in lectures.	4.32	1.05	0.70
2- Instructor utilizes lecture time efficiently.	4.29	1.05	0.75
3- Instructor speaks with clear and audible voice.	4.27	1.07	0.69
4- Instructor prepares material well.	4.23	1.07	0.65
5- Instructor distributes syllabus and explains it.	4.16	1.12	0.76
6- Instructor showed broad knowledge in material.	4.15	1.10	0.74
7- Instructor encourages students to ask questions and give opinions whenever needed.	4.11	1.16	0.78
8- Instructor gives enough time for queries during office hours.	4.04	1.11	0.82
9- Course objectives have been explained clearly.	4.04	1.15	0.66
10- Course objectives have been achieved.	4.00	1.14	0.84
11-Instructor evaluates students with justice.	3.95	1.21	0.68
12- Instructor's homework and exams conform to material covered in the course.	3.93	1.27	0.85
13- Instructor makes learning easy and interesting.	3.84	1.29	0.79
14- I would like to study other courses with this instructor.	3.82	1.44	0.81
15- This course requires time and effort matching its credit hours.	3.96	1.19	0.69
16- I recommend my colleagues to study this course.	3.90	1.27	0.38
17- Textbook and supporting studying materials are suitable.	3.77	1.26	0.71
18- General classroom surrounding (temperature, lighting, seats, equipment…) are appropriate.	3.49	1.39	0.56

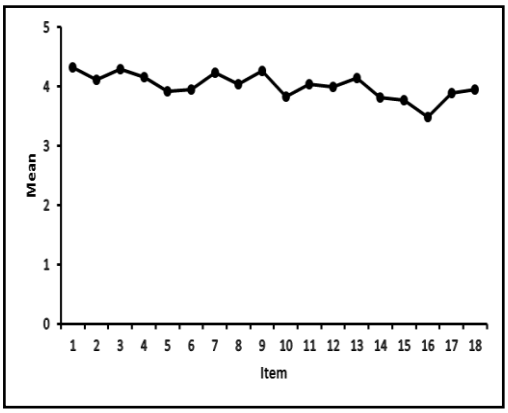

Figure 1. Students' mean ratings for the SET items.

Exploratory Factor Analysis

To determine the dimensionality and the factor structure of the developed scale, the 18 items were analyzed using principal component analysis with direct oblimin rotation. The value of Kaiser-Meyer-Olkin measure of sampling adequacy (Mulaik, 2010) of 0.98 showed that data is meritorious for factor analysis. The large value of (χ^2 = ,*78273.62, p* < .001) showed that correlation matrix is not an identity matrix and variables are positively correlated with each other. Two components were retained with eigenvalues over Kaiser's criterion of 1 (Mulaik, 2010) that explained 63.14% of the total variance. The correlation coefficient between the two dimensions was 0.53, which justifies the use of oblique rotation. Table 2 shows the factor loadings after rotation.

Table 2 shows that 14 items clustered on the same factor that had an eigenvalue of 10.27 and accounted for 57.04% of the total variance. This factor was related to the instructor and teaching effectiveness, while the second factor with the last four items was related to the course attributes which explained 6.1% of the total variance.

Initial Estimates of Reliability

Internal consistency estimates of reliability (Cronbach's alpha) was computed for the whole instrument and for the two dimensions, given that the two factors were correlated. The whole instrument had high internal consistency; Cronbach's alpha was 0.94. Regarding the first factor that relates to the instructor's ability, Cronbach's alpha was 0.94. However, for the second factor Cronbach's alpha was 0.65.

Table 2
Factor loadings of the 18 Items of the Student Evaluation of Teaching Scale through Principal Component Analysis Using Direct Oblimin Method

	Item	Factor I	Factor II
1.	Mutual respect predominates in lectures.	0.83	
2.	Instructor encourages students to ask questions and give opinions whenever needed.	0.84	
3.	Instructor utilizes lecture time efficiently.	0.86	
4.	Instructor distributes syllabus and explains it.	0.76	
5.	Instructor's homework and exams conform to material covered in the course.	0.67	
6.	Instructor evaluates students with justice.	0.61	
7.	Instructor prepares material well.	0.89	
8.	Course objectives have been explained clearly.	0.82	
9.	Instructor speaks with clear and audible voice.	0.73	
10.	Instructor makes learning easy and interesting.	0.77	
11.	Instructor gives enough time for queries during office hours.	0.62	
12.	Course objectives have been achieved.	0.75	
13.	Instructor showed broad knowledge in material.	0.82	
14.	I would like to study other courses with this instructor.	0.74	
15.	Textbook and supporting studying materials are suitable.		0.60
16.	General classroom surrounding (temperature, lighting, seats, equipment…) are appropriate.		0.76
17.	I recommend my colleagues to study this course.		0.53
18.	This course requires time and effort matching its credit hours.		0.64
	Eigenvalue	10.27	1.10
	% of variance	57.٠4	6.10
	Cumulative % of variance	57.٠4	63.14

Confirmatory Factor Analysis

In the structural equation model, there were 18 observed variables and two latent variables. The observed variables were the instrument items, while the latent variables represent the two factors extracted using exploratory factor analysis, i.e., teacher or instructor ability and course attributes.

CFA was conducted to determine whether each of the 18 observed variables appropriately loaded on the two latent variables. Figure 2 shows the path diagram for the model. As observed in Figure 2, all items had significant factor loadings ($p < 0.05$) on the two latent variables, ranging from 0.68 to 0.87 on teacher ability and from 0.44 to 0.82 on course attributes.

In the model, $\chi^2 = 4352.93$, df =134, $p < 0.001$. The large and significant χ^2 value indicated there was a significant difference between sample and population covariance matrix, which showed that this model was a poor fit to the data. However, given the large sample in this study, it is not surprising that the χ^2 is significant. However, the values of the other fit indices show a good fit to the data. In this model, RMSEA = 0.072 (which was below 0.10), SRMR = 0.029 (which was below 0.05), and GFI = 0.93 (which was above 0.90).

Convergent Validity

As an indicator of convergent validity, Pearson correlation between course averages of the first factor that is related to the teacher effectiveness and course averages of students' final grades in each course was 0.30, $p < 0.001$. This significant positive correlation between SETs and examination scores at the course level suggests that effective teaching affects both, student ratings and examination results, which supports the validity of SETs as measures of teaching effectiveness.

Reliability

Finally, the whole instrument had high internal consistency; Cronbach's alpha was 0.95. Regarding the first factor that relates to the instructor's ability, Cronbach's alpha was 0.96. However, for the second factor Cronbach's alpha was 0.76.

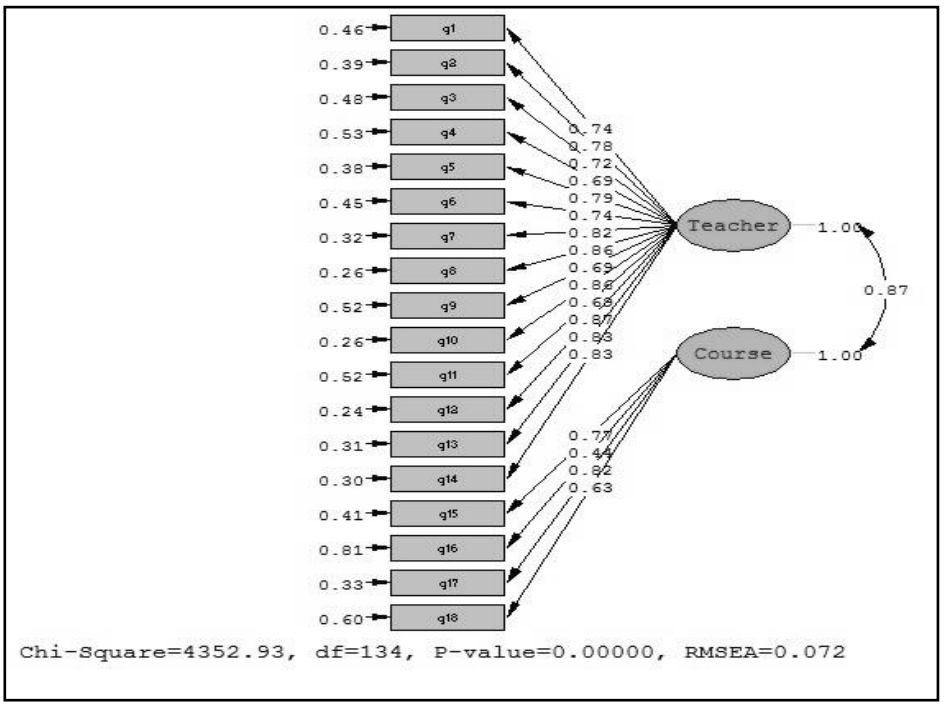

Figure 2: Path diagram for student evaluation of instruction instrument

Discussion and Conclusions

Many universities have developed their own instruments to be used in making judgments about instructors teaching effectiveness. The purpose of this study was to investigate the psychometric properties (i.e., validity and reliability) of an instrument developed at a major public Jordanian university to assess student evaluation of teaching effectiveness. In the current study, validity was investigated using evidences from internal structure and convergent validity. On the other hand, Cronbach's alpha was used an estimate of reliability.

Regarding the construct validity of the questionnaire, the results of the principal component analysis revealed that a two-factor solution accounted for 63.١4% of the total variation in the students' responses to the 18-item questionnaire. However, the part of the questionnaire that relates directly to the instructor approved to be unidimensional while accounting for 57.٠٤% of the total variance. Moreover, the results of CFA supported this finding in that the first 14 items captured teaching effectiveness. These results are consistent with previous research that approved a one-factor solution (Bassi et al., 2017; Catano & Harvey, 2011; Chen & Watkins 2010; Oon et al., 2016; Zhao & Gallant, 2012).

On the other hand, this study showed inconsistency in findings of factor structure with many previous studies that resulted in a multi-factor model for SET (Nemec et al., 2018). This can be explained by the existing disagreement in the literature on whether the different dimensions are discrete or are representative of a single higher-order teaching effectiveness dimension (Patrick & Smart, 1998). Another possible reason that might lead to the one-factor solution is that the instrument consisted of a small number of items such that each dimension of effective teaching was captured by one or two items.

Convergent validity was assessed by examining the relationship of course averages of SET scores on the first factor to course averages of student achievement (final grades in the course). The positive significant correlation of 0.30 revealed that effective teaching affects both, student ratings and examination results, which supports the validity of SETs as measures of teaching effectiveness. This is consistent with the findings of other studies that reported a correlation coefficient varying between 0.10 and 0.47 (e.g., Catano & Harvey, 2011; Onwuegbuzie et al., 2007)

In this study, reliability was estimated by internal-consistency reliability indices. Cronbach's alpha for the first factor that relates to the instructor ability was 0.96 indicating that only about 4% of the variability in the students' responses were attributed to error. This finding resonates with the finding of other studies that used internal consistency indices in estimating reliability (Bassi et al., 2017; Catano & Harvey, 2011; Chen & Watkins 2010; Nemec et al., 2018; Oon et al., 2016; Zhao & Gallant, 2012). These studies reported high levels of reliability (>0.90) when estimated using Cronbach's alpha.

Although the findings of this study provided validity and reliability support for the instrument, caution should be exercised in the interpretation of the data. High mean ratings on almost all items reflects a possible ceiling effect in students' ratings of teaching effectiveness. One possible reason for this is that many instructors noticed the impact of SET scores on their career given that HU used the results of SET for summative purposes. No faculty member can apply for promotion if his/her score were low for three semesters in a row. This might lead them to practices aimed at increasing their SET scores rather than improving their teaching (Simpson & Siguaw, 2000).

It is possible that factors unrelated to teaching effectiveness but related to student characteristics, lecturer behaviour, and the course administration could have attributed to the high mean ratings (d'Apollonia & Abrami, 1997). Shevlin, Banyard, Davies, and Griffiths (2000) argued that if students have a positive personal and/or social view of the lecturer, this may lead to more positive ratings irrespective of the actual level of teaching effectiveness. In this study, mutual respect predominated in lectures as reflected in

the highest mean ratings. This might have affected students' ratings to be positive on all other items irrespective of the actual performance of the instructor.

Moreover, grading leniency might have led to the high mean students' rating. It is one of the instructor-related variables, which showed to have a strong positive relationship with ratings of teaching effectiveness (Greenwald & Gillmore, 1997). Finally, another factor that possibly contributed to the high mean ratings is class size. Large classes and small classes tended to give the most positive ratings (Benton & Cashin, 2014).

The results of the current study provided evidence to support validity and reliability of the instrument. However, content validity was not supported in the current study. It is important to know what was the theoretical framework that underlies the development of this instrument. It is also of importance to check the content of the instrument for representativeness and relatedness. Moreover, this instrument lacks more evidences of convergent validity through correlating the scores obtained using this instrument and the scores obtained using other well-known measures of teaching effectiveness.

Benton & Cashin (2014) recommended the use of multiple sources of data in making a valid judgment about an instructor's overall teaching effectiveness. They asserted that no single source of information—including student ratings—provides such sufficient evidence. Moreover, there are important aspects of teaching that students are not competent to rate. For example, subject-matter knowledge, course design, curriculum development, and commitment to teaching. Therefore, caution must be taken when using this instrument in evaluating faculty members and making promoting decisions.

One limitation in this study was that it used only two samples in two independent semesters. Further research in SET that is related to the HU is needed. For example, this instrument needs to be revalidated using data from subsequent semesters to check if the one-factor model is consistent across different samples. Moreover, research is needed to compare the results of SET obtained at the HU with those obtained at other universities, especially in the west. Moreover, to investigate the effect of SET on the teaching quality of instructors, more research is needed in comparing the findings of SET for the same instructor across different semesters. Finally, it is a good practice to incorporate the demographic variables into the research by studying its effect on the mean rating scores. One also could study differential item functioning according to gender, year, major, etc. as another evidence of validity.

References

Aleamoni, L. M. (1978). Development and factorial validation of the Arizona course/instructor evaluation questionnaire. *Educational and Psychological Measurement, 38*(4), 1063-1067. doi:10.1177/001316447803800426.

Balam, E. M., & Shannon, D. M. (2009). Student ratings of college teaching: A comparison of faculty and their students. *Assessment & Evaluation in Higher Education, 35*(2), 209-221. doi:10.1080/02602930902795901.

Bassi, F., Clerci, R., & Aquario, D. (2017). Students' evaluation of teaching at a large Italian university: validation of measurement scale. *Electronic Journal of Applied Statistical Analysis, 10*(1), 93-117. Doi: 10.1285/i20705948v10n1p93.

Benton S.& Cashin W. (2014) Student Ratings of Instruction in College and University Courses. In: Paulsen M. (eds) Higher Education: Handbook of Theory and Research (vol. 29, pp. 279-326). Dordrecht: Springer,. https://doi.org/10.1007/978-94-017-8005-6_7.

Catano, V. M., & Harvey, S. (2011). Student perception of teaching effectiveness: development and validation of the Evaluation of Teaching Competencies Scale (ETCS). *Assessment & Evaluation in Higher*

Education, 36(6), 701-717. doi:10.1080/02602938.2010.484879.

Chen, G. H., & Watkins, D. (2010). Stability and correlates of student evaluations of teaching at a Chinese university. *Assessment & Evaluation in Higher Education, 35*(6), 675-685. doi:10.1080/02602930902977715.

d'Apollonia, S., & Abrami, P. C. (1997). Navigating student ratings of instruction. *American Psychologist, 52*(11), 1198-1208. doi:10.1037/0003-066X.52.11.1198.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London: Sage Publications Ltd.

Greenwald, A. G., & Gillmore, G. M. (1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist, 52*(11), 1209-1217. doi:http://dx.doi.org/10.1037/0003-066X.52.11.1209.

Griffin, B. W. (2001). Instructor reputation and student ratings of instruction. *Contemporary Educational Psychology, 26*(4), 534-552. doi:http://dx.doi.org/10.1006/ceps.2000.1075.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 65-110). Westport, CT: American Council on Education and Praeger Publishers.

Jöreskog, K., & Sörbom, D. (2006). LISREL 8.80. Lincolnwood, IL: Scientific Software International, Inc.

Mulaik, S. (2010). *Foundations of facotr analysis* (2nd ed.). FL:CPC Press.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17-64). Westport, CT: American Council on Education and Praeger Publishers.

Kelloway, K. E. (1998). *Using LISREL for structural equation modelling: A researcher's guide.* Thousand Oaks, CA: Sage.

Kogan, L. R., Schoenfeld-Tacher, R., & Hellyer, P. W. (2010). Student evaluations of teaching: Perceptions of faculty based on gender, position, and rank. *Teaching in Higher Education, 15*(6), 623-636. doi:10.1080/13562517.2010.491911.

Lemos, M. S., Queirós, C., Teixeira, P. M., & Menezes, I. (2011). Development and validation of a theoretically based, multidimensional questionnaire of student evaluation of university teaching. *Assessment & Evaluation in Higher Education, 36*(7), 843-864. doi:10.1080/02602938.2010.493969.

Marsh, H. W. (1982). SEEQ: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology, 52*(1), 77-95. doi:10.1111/j.2044-8279.1982.tb02505.x.

Martínez-Gómez, M., Sierra, J. M. C., Jabaloyes, J., & Zarzo, M. (2010). A multivariate method for analyzing and improving the use of student evaluation of teaching questionnaires: A case study. *Quality & Quantity, 45*(6), 1415-1427. doi:10.1007/s11135-010-9345-5

Morley, D. (2014). Assessing the reliability of student evaluations of teaching: Choosing the right coefficient. *Assessment & Evaluation in Higher Education, 39*(2), 127-139. doi:10.1080/02602938.2013.796508.

Nemec, E. C., Baker, D. M., Zhang, D., & Dintzner, M. (2018). Development of valid and reliable tools for student evaluation of teaching. *Currents in Pharmacy Teaching and Learning, 10*(5), 549-557. doi:10.1016/j.cptl.2018.02.009.

Nunnally, J., & Bernstein, I. (1994). *Psychometric theory* (3rd ed.). NewYork: McGraw-Hill.

Onwuegbuzie, A. J., Daniel, L. G., & Collins, K. M. T. (2007). A meta-validation model for assessing the score-validity of student teaching evaluations. *Quality & Quantity, 43*(2), 197-209. doi:10.1007/s11135-007-9112-4.

Oon, P.-T., Spencer, B., & Kam, C. C. S. (2016). Psychometric quality of a

student evaluation of teaching survey in higher education. *Assessment & Evaluation in Higher Education, 42*(5), 788-800. doi:10.1080/02602938.2016.1193119.

Ory, J. C. (2001). Faculty thoughts and concerns about student ratings. *New Directions for Teaching and Learning, 2001*(87), 3-15. doi:10.1002/tl.23.

Patrick, J., & Smart, R. M. (1998). An empirical evaluation of teacher effectiveness: The emergence of three critical factors. *Assessment & Evaluation in Higher Education, 23*(2), 165-178. doi:10.1080/0260293980230205.

Penny, A. R. (2003). Changing the agenda for research into students' views about university teaching: Four shortcomings of SRT research. *Teaching in Higher Education, 8*(3), 399-411. doi:10.1080/13562510309396.

Ramsden, P. (1991). A performance indicator of teaching quality in higher education: The course experience questionnaire. *Studies in Higher Education, 16*(2), 129-150. doi:10.1080/03075079112331382944.

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

Schumacker, R. E., & Lomax, R. G. (2010). *A beginner's guide to structural equation modeling* (3rd ed.). New York: Routledge.

Shevlin, M., Banyard, P., Davies, M., & Griffiths, M. (2000). The validity of student evaluation of teaching in higher education: Love me, love my lectures? *Assessment & Evaluation in Higher Education, 25*(4), 397-405. doi:10.1080/713611436.

Simpson, P. M., & Siguaw, J. A. (2000). Student evaluations of teaching: An exploratory study of the faculty response. *Journal of Marketing Education, 22*(3), 199-213. doi:10.1177/0273475300223004.

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. doi:10.3102/0034654313496870.

Stehle, S., Spinath, B., & Kadmon, M. (2012). Measuring teaching effectiveness: Correspondence between students' evaluations of teaching and different measures of student learning. *Research in Higher Education, 53*(8), 888-904. doi:10.1007/s11162-012-9260-9.

Toland, M. D., & De Ayala, R. J. (2005). A multilevel factor analysis of students' evaluations of teaching. *Educational and Psychological Measurement, 65*(2), 272-296. doi:10.1177/0013164404268667.

Willits, F. & Brennan, M. (2017). Another look at college student's ratings of course quality: data from Penn State student surveys in three settings. *Assessment & Evaluation in Higher Education, 42*(3), 443-462. DOI: http://dx.doi.org/10.1080/02602938.2015.1120858.

Wilson, K. L., Lizzio, A., & Ramsden, P. (1997). The development, validation and application of the course experience questionnaire. *Studies in Higher Education, 22*(1), 33-53. doi:10.1080/03075079712331381121.

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. doi:10.1080/02602938.2010.523819.