The Validity and Reliability of Student Evaluation of Teaching at the National University of Lesotho

*Peter Khaola and Regina Thetsane**

Abstract
Many higher education institutions use the Student Evaluation of Teaching (SET) scale to evaluate the quality of instructors' teaching. It includes students' evaluation of the teacher, the teaching process, teaching approaches and the learning outcomes. Due to its reported dubious reliability and validity, and inherent bias in measuring the quality of teaching, SET remains a hotly debated and controversial instrument. This study evaluated the reliability and validity of the SET scale adopted by the National University of Lesotho. Self-administered SET questionnaires were distributed to 104 third- and fourth-year Bachelor of Commerce students to evaluate ten lecturers, resulting in 751 assessment records. The data were analysed using the Statistical Package for the Social Sciences (SPSS) and Partial Least Squares Structural Equation Modelling (PLS-SEM). While the findings suggest that the SET instrument used at the university is reasonably reliable and valid, minor concerns were raised with regard to discriminant validity, and serious concerns in relation to content validity. Based on the existing literature and the psychometric properties of this SET instrument, it is recommended that university management exercise caution in using its results to make evaluative personnel decisions such as promotions, confirmations, and dismissals. It is also recommended that the SET instrument should be revised and validated and be primarily used for formative purposes such as obtaining feedback for the development of individual instructors.

Key words: formative assessment, reliability, student evaluation of teaching, summative assessment, validity

ABOUT THE AUTHORS: PETER KHAOLA AND REGINA THETSANE*, National University of Lesotho. Email: peterkhaola@gmail.com, makoloithetsane@gmail.com
* Corresponding author

De nombreux établissements d'enseignement supérieur utilisent l'Echelle de l'évaluation de l'enseignement par les étudiants (SET), pour évaluer la qualité d'enseignement des instructeurs. Ledit procédé comprend l'évaluation, par les étudiants, de l'enseignant, du processus d'enseignement, des approches pédagogiques et des résultats d'apprentissage. En raison de sa douteuse fiabilité et validité signalées, ainsi que de son parti pris inhérent en matière de jugement de la qualité de l'enseignement, l'échelle SET reste un outil très débattu et controversé. Cette étude a évalué la fiabilité et la validité de l'échelle SET adoptée par l'Université nationale du Lesotho. Des questionnaires SET auto-administrés ont été distribués à 104 étudiants de troisième et de quatrième années du baccalauréat en commerce afin d'évaluer dix chargés de cours, ce qui a donné lieu à 751 dossiers d'évaluation. Les données ont été analysées à l'aide du Paquet statistique pour les sciences sociales (SPSS) et de la Modélisation partielle des équations structurelles des moins carrés (PLS-SEM). Bien que les conclusions donnent à penser que l'outil SET utilisé à l'université est raisonnablement fiable et valide, des préoccupations mineures ont été soulevées au sujet de la validité discriminatoire et de graves préoccupations concernant la validité du contenu. D'après la documentation existante et les propriétés psychométriques de cet outil SET, il est recommandé que la direction de l'université fasse preuve de prudence dans l'utilisation de ses résultats pour prendre des décisions évaluatives concernant le personnel, comme les promotions, les confirmations et les congédiements. Il est également recommandé que l'outil SET soit révisé et validé et qu'il soit d'abord utilisé à des fins de formation, comme l'obtention de commentaires pour le développement des instructeurs individuels.

Mots-clés: évaluation formative, fiabilité, évaluation de l'enseignement par les étudiants, évaluation sommative, validité

1. Background and Introduction

In recent years, Lesotho's higher education sector has undergone significant change and substantial growth. While the National University of Lesotho (NUL) was the sole university in the country from 1975 to 2008, the founding of two private universities introduced a modicum of choice for students and triggered competition. This calls for Higher Education Institutions (HEIs) to differentiate themselves by improving the quality of their teaching and to evaluate teaching competence through the use of student evaluation of teaching (SET) (Marks, 2000). For the purposes of this article, SET is a general term used to describe the process of using student input on their teachers' overall activities and attitudes. It involves

evaluating the teacher, the teaching process, teaching approaches and learning outcomes as perceived by students.

According to Clayson (2009) and Morley (2014), the first published article on evaluations was produced by researchers at Purdue University in the 1920s. In the early 1950s, the University of Washington became one of the first institutions to conduct formal evaluations. Since the 1970s, SET has been used almost universally by HEIs, especially in Western countries (Clayson, 2009; Linse, 2017; Morley, 2014; Spooren, Brockx, and Mortelmans, 2013; Uttl, White and Gonzalez, 2017).

There is however, a paucity of research on student evaluations in Africa (Tennant and Khamis, 2017) and SET was only recently introduced at NUL, despite the fact that the institution's promotion criteria require lecturers to demonstrate competence in research, teaching and community service. Its adoption may have primarily been influenced by the National Council on Higher Education (CHE, established by section 4 of the Higher Education Act of 2004) which includes such ratings in the list of standards it employs to accredit tertiary institutions and their programmes in Lesotho.

Despite concerns pertaining to its reliability and validity (Penny, 2003), SET is likely to continue to be employed by HEIs (Hornstein, 2017; Linse, 2017) due to a number of reasons. First, SET is valued by both students and administrators as a cost-effective tool that also gives students some voice as consumers of higher education (Hornstein, 2017; Spooren et al., 2013). Second, the use of SET for purposes of teaching improvement (formative purposes) is widely supported (Penny, 2003). Third, there is evidence (albeit subdued) that SET can improve students' learning (Clayson, 2009; Cohen, 1981; Spooren et al., 2013). Finally, it contributes to the evaluation of teaching effectiveness in making decisions pertaining to personnel, including promotions, contract renewal and tenure (Cagri, 2017).

We argue that, rather than focussing on whether or not SET should be discontinued, HEIs could benefit from research that focuses on how best to use this tool, including how to design, develop and validate SET scales, and how to address unresolved issues.

Objectives of the Study

Our main purpose was to evaluate the scale used by students to rate the teaching effectiveness of NUL lecturers. The evaluation focussed on the third of the three phases of scale development and validation recommended by Boateng, Neilands, Frongillo, Melgar-Quinonez and Young (2018). According to Boateng et al. (2018), the three phases and nine steps of scale development and validation are: a) *item development* (identification of domain and item generation and content validity), b) *scale development* (pretesting of questions, sampling and survey administration, item reduction,

and extraction of factors), and c) *scale evaluation* (tests of dimensionality, tests of reliability, and tests of validity). The study focused on scale evaluation because the SET scale already exists at NUL.

The secondary purpose of the study was to provide recommendations based on the reviewed literature, and the findings of the scale evaluation exercise.

The study was guided by the following broad research questions:

1) Is the SET instrument used at NUL reliable and valid?
2) What lessons can be learned from the literature and the SET instrument used at NUL?

2. Literature Review

2.1 Student Evaluation of Teaching

Student evaluation of teaching is one of the methods used by educational institutions to assess the effectiveness of teaching (Little, Goe and Bell, 2009). While there is no universal definition of the term 'effectiveness of teaching', it has been defined narrowly as the lecturer's ability to improve students' learning (as measured through students' grades); and broadly as the lecturer's ability to impart wide ranging skills that shape students to be better citizens within and outside the classroom (Little et al., 2009). We adopted the broader definition.

Student evaluation of teaching is based on the widely accepted axiom that students learn effectively if they are taught by highly rated lecturers (Uttl et al., 2017), and not necessarily by highly qualified ones (Little et al., 2009). Put differently, SET is predicated on the realisation that recruitment of highly qualified teachers is a necessary, but not a sufficient condition for students' success in higher education.

The typical scales used for SET have four to five points ranging from strongly disagree to strongly agree (Uttl et al., 2017; McBean and Al-Nassri, 1982). They evaluate factors such as course content, the lecturer's course knowledge, clarity of explanation, preparation for lectures, enthusiasm for the course, fairness in marking, friendliness, availability, approachability, etc. (Uttl et al., 2017).

2.2 Purpose of SET

Although SET was originally designed and used for formative purposes (providing feedback for improvement of lecturers), it has also been used for other purposes, including summative purposes (providing input for personnel decisions such as merit pay, tenure, promotions, and dismissals), and for demonstrating institutional accountability in ensuring the quality of the education provided (Spooren et al., 2013).

The instrument's popularity derives from the ease and cost-effectiveness it offers in collecting, presenting and interpreting data (Hornstein, 2017; Spooren et al., 2013). It has been argued that student ratings are both cost and time-efficient, and require minimal training (Little et al., 2009). Student Evaluation of Teaching also resonates well with the new principles of managerialism (market principles) in education in which students are considered bona fide customers, and teachers are considered service providers (Hornstein, 2017). It allows students to express their class-room experiences and levels of satisfaction as valued customers (Uttl et al., 2017), with these sometimes likened to the quality of teaching and teacher performance (Spooren et al., 2013). While this is problematic because 'liking' and 'learning' are two different concepts, it is generally agreed that HEIs should seek and obtain student satisfaction. However, the use of SET has been the subject of drawn-out debate among researchers and educators (Hornstein, 2017; Spooren et al., 2013).

2.3 The Reliability and Validity of SET

Reliability assesses the degree to which the scale accurately provides consistent measures, while validity refers to the extent to which the scale measures what it purports to measure (Cooper and Schindler, 2014).

Some analysts argue that it is difficult to define 'effectiveness of teaching', calling into question the validity and usefulness of SET as a measure of faculty competence (Hornstein, 2017). Opponents of SET further argue that it is a measure of student satisfaction, and so to speak, a 'popularity contest' (Hornstein, 2017; Uttl et al., 2017) which does not reflect lecturers' competence (Clayson, 2009) and capability of delivering quality teaching (Hornstein, 2017). Spooren et al. (2013) assert that if SET is used primarily for summative or administrative purposes, it can encourage lecturers to engage in practices that aim at increasing SET scores at the expense of quality of teaching. For instance, it could lead to 'grading leniency' and 'grade inflation' that is totally unrelated to the acquisition of knowledge (Spooren et al., 2013).

While these concerns are valid, they are by no means insurmountable. For instance, researchers could expend more effort on conceptualisation (definition) of SET, and administrators could be advised not to use it as the only measure of effectiveness (Little et al., 2009). Students can also be trained to understand the purpose and use of SET in universities.

It has also been argued that students may not have the ability to assess the quality of the curriculum and the lecturer's content knowledge (Little et al., 2009). While some studies report low correlation between student learning and SET (Clayson, 2009), others conclude that there is an insignificant relationship between the two constructs (Boring, Ottoboni, and Stark, 2016; Uttl et al., 2017). Students' ratings are also said to be suscep-

tible to leniency, bias and halo error (Little et al., 2009). Several studies have found that SET is consistently biased against female lecturers who are evaluated more negatively than their male counterparts (Boring et al., 2016; Mitchell and Martin, 2018; MacNell, Driscoll and Hunt, 2015).

These concerns notwithstanding, students remain best-positioned to evaluate certain elements of teaching (e.g., the lecturer's enthusiasm, friendliness and availability). It is also important to remember that bias is not only encountered in SET; the same biases exist in other employment-related issues, including promotions, salaries, and performance appraisals (Linse, 2017). Linse (2017) argues that these biases not only fail to fully explain the consistently low ratings of some lecturers, but are also not widespread.

In terms of reliability, Morley's (2014) review of the existing literature suggests that SETs are reliable measures of teaching effectiveness. However, since reliability is a necessary, but not a sufficient condition for validity, the use of SET continues to evoke mixed feelings among faculty members.

Another criticism of SET is that, in some cases, its scales have not been validated with regard to their psychometric properties (Spooren et al., 2013). For instance, this article is based on the SET scale at NUL which is yet to be validated. While this is problematic because the university envisages using the results for a variety of purposes (e.g., promotions, pay increases, and training and development), arguably, SET scales can be evaluated and improved.

In summary, while there is on-going debate on the validity and reliability of SET, its use is set to continue in HEIs (Linse, 2017), mainly because, if well-designed, it can be a reasonably valid instrument to assess teaching quality (Spooren et al., 2013). However, users should be made aware of its weaknesses and biases, and be advised to apply it judiciously.

3. Method

The SET instrument used in the study was developed by NUL's Human Resource Department and the Centre for Teaching and Learning (CTL), and approved by the Senate and Council. Formal SET questionnaires were distributed to 104 third- and fourth-year students enrolled in a B.Com. programme at NUL to evaluate ten members of staff, giving rise to 751 assessment records (about 75 evaluations per lecturer). Four staff members are female, and six are male.

Students were informed that participation in the study was voluntary, and that they could withdraw at any time without fear of negative consequences. Furthermore, no rewards were promised or given to students for participating in the study.

The survey was undertaken towards the end of the second semester, by which time students are familiar with their instructors. Students in their final years of study enrolled in a business programme were targeted for two main reasons. First, researchers recommend evaluation based on students from similar programmes (Clayson, 2009). Second, compared to freshmen and sophomores, these students are mature and knowledgeable, and arguably should provide more reliable data. Thus, the sampling approach was used to allay fears that (new) students may not be qualified to assess their lecturers (Hornstein, 2017; Little et al., 2009; Spooren et al., 2013).

The SET scale used at NUL does not assess students' gender or age, but the participants were asked to rate the teaching capability of six male and four female lecturers.

The data were analysed using the Statistical Package for the Social Sciences (SPSS) and the partial least squares structural equation modelling (PLS-SEM, Smart PLS 3).

Unless otherwise indicated, items were measured on a Likert-scale ranging from 1 (strongly disagree) to 5 (strongly agree).

4. Results and Discussion

4.1 Dimensionality of SET

Exploratory factor analysis (EFA) was used to examine the dimensionality of the SET instrument under review. The results are shown in Table 1.

The Kaizer-Meyer-Olkin (KMO) measure of sampling adequacy test indicated an adequate figure of 0.92 (which is better than the threshold of 0.6), and the Bartlett's test of sphericity ($X^2 = (153) = 4141.322$, $p \leq 0.001$) was significant and hence acceptable. Sphericity and KMO tests are often used to examine the degree to which factorisation is applicable and suitable to responses to items (Siebert and Kunz, 2016).

After deleting the item ('the lecturer returned students' work in a timely manner') that loaded ambiguously on more than one factor, three dimensions (explaining 56.829% of variance) emerged from EFA. Factor 1 tapped into course delivery; factor 2 into course assessment and support; and factor 3 into lecturer attendance. Although there is no consensus on the number of dimensions of SET, the multi-dimensional nature of the SET instrument at NUL is in line with many instruments used in the literature. For instance, in their review of the validity of SET instruments, Spooren et al. (2013) found that dimensions of popular instruments range from two to 12. Multiple dimensions of SET are said to be justified because good teaching is reflected in multiple aspects (Spooren et al. 2013, p. 607).

Table 1. Exploratory Factor Analysis Results

	Factors		
	1	2	3
The lecturer's explanations were clear and practical	.832	.244	.105
The lecturer demonstrated knowledge of this course	.810	.259	.123
The lecturer came well prepared for each class	.781	.134	.248
Class sessions were well organised	.774	.190	.208
The lecturer demonstrated enthusiasm for teaching this course	.561	.353	.261
The lecturer stimulated my interest in this course	.526	.249	-.018
The lecturer made assessment requirements clear	.272	**.706**	.067
The lecturer used appropriate and fair assessment methods	.292	**.683**	.099
Students' work was returned with useful, constructive feedback	.245	**.645**	.109
The lecturer set high standards of achievement	.299	**.606**	.174
The lecturer encouraged participation and independent thinking	.342	**.605**	.177
The lecturer was readily available to students	.131	**.587**	.343
The lecturer seemed sensitive to and concerned about students' progress	-.047	**.527**	.194
The lecturer used interactive methods of teaching	.246	**.423**	-.148
The lecturer was punctual for class sessions	.073	.146	**.799**
The lecturer's attendance was good	.218	.134	**.755**
The lecturer gave course outlines with clear course objectives and outcomes	.148	.126	**.598**
Variance explained (%)	40.167	9.075	7.136

Extraction Method: Principal Component Analysis. **Rotation Method:** Varimax with Kaiser Normalization. Rotation converged in 6 iterations.

A closer look at Table 1 suggests that there was some overlap of related items. For instance, factor 1 consists of the lecturer's course knowledge and enthusiasm in the delivery of course material; factor 2 consists of assessment, feedback and the helpfulness of the course instructor; and factor 3 mainly relates to the lecturer's attendance. It could also be argued that the two items that fall under factor 2, namely, 'the lecturer encouraged participation and independent thinking', and 'the lecturer used interactive methods for teaching' relate more to factor 1 than 2, but a more reflective and closer look suggests that these items also capture some elements of

formative assessment because interactive teaching and student participation allow the lecturer to provide formative feedback to learners.

In summary, in line with the SET instruments in the literature, the SET instrument used at NUL is multi-dimensional. This reflects the multiple aspects of good teaching. However, the multiplicity of SET instruments used in the literature is a source of concern for theory testing, validation and improvement.

4.2 The Normal Distribution of SET Items

Concerns have been expressed that SET items are ordinal and skewed (i.e., not normally distributed) (Linse, 2017). Tests for skewness and kurtosis were used to examine the normal distribution of the items in SET. Skewness refers to the asymmetry of the distribution or its departure from symmetry (George and Mallery, 2010). The rule of thumb suggests that if skewness is greater than [±1], the distribution is considered highly skewed. Another measure used to measure normality is kurtosis, which refers to the 'peakedness' of the distribution (George and Mallery, 2010). The rule of thumb is that kurtosis of ±1.96 for small samples deviates problematically from normal distribution. Like skewness, the greater the kurtosis in absolute terms, the more the distribution deviates from normal distribution.

The results of the skewness and kurtosis of the SET items are presented in Table 2.

Of the 18 items used in the SET scale, only four deviated problematically from normal distribution. As shown in Table 2, these items are 'the lecturer stimulated my interest in this course'; 'the lecturer seemed sensitive and concerned about students' progress', 'the lecturer used interactive methods for teaching', and 'the lecturer gave course outlines with clear course objectives and outcomes'. The first three items were positively skewed, suggesting that the ratings tended to cluster at the lower end of the scale. The last item was negatively skewed, suggesting that the ratings tended to cluster at the higher end of the scale. This difference may have been caused by the fact that lecturers at NUL use a standardised template for course outlines, explaining why the majority scored higher on provision of course outlines with clear objectives and learning outcomes.

Linse (2017) asserts that, student rating distributions are typically negatively skewed (i.e., not normally distributed) and tend to cluster at the high end of the scale. This claim is supported in this case because of 18 items, only three were positively skewed.

In summary, the distribution of most of the SET items used at NUL does not differ problematically from normal distribution, and largely tends to cluster at the higher end of the SET scale.

Table 2. Descriptive Statistics and Measures of Skewness and Kurtosis

	Median	Mean	SD	Skewness	Kurtosis
The lecturer's explanations were clear and practical	4.00	3.85	1.21	-0.79	0.34
The lecturer demonstrated knowledge of this course	4.00	4.17	1.00	-1.22	1.08
The lecturer came well prepared for each class	5.00	4.24	0.93	-1.18	1.00
Class sessions were well organised	4.00	4.10	1.00	-1.01	0.43
The lecturer demonstrated enthusiasm for teaching this course	4.00	4.09	0.95	-1.08	1.09
The lecturer stimulated my interest in this course	4.00	3.67	1.99	10.95	225.20
The lecturer made assessment requirements clear	4.00	3.86	1.02	-0.72	-0.02
The lecturer used appropriate and fair assessment methods	4.00	3.81	1.00	-0.61	-0.11
Students' work was returned with useful, constructive feedback	4.00	3.51	1.21	-0.39	-0.76
The lecturer set high standards of achievement	4.00	3.79	0.97	-0.53	-0.03
The lecturer encouraged participation and independent thinking	4.00	4.09	1.00	-1.08	0.83
The lecturer was readily available to students	4.00	4.15	0.90	-0.97	0.75
The lecturer seemed sensitive to and concerned about students' progress	4.00	3.87	2.25	16.68	389.15
The lecturer used interactive methods of teaching	4.00	3.88	2.11	17.73	422.17
The lecturer was punctual for class sessions	4.00	4.20	0.94	-1.15	0.99
The lecturer's attendance was good	5.00	4.44	0.76	-1.37	1.78
The lecturer gave course outlines with clear course objectives and outcomes	5.00	4.76	0.55	-2.86	10.46
Deleted Item					
The lecturer returned students' work in a timely manner	4.00	3.81	1.24	-0.78	-0.41

4.3 The Reliability of the SET Scale

As noted previously, reliability assesses the degree to which the scale accurately provides consistent measures (Cooper and Schindler, 2014; Hair, Black, Babin and Anderson, 2010). For SET to be accepted by instructors, it must produce consistent results when used by different students.

To measure the internal reliability or consistency of a scale, among other measures, researchers often calculate the Cronbach's alpha or composite reliability (Hair, Ringle and Sarstedt, 2011; Hair, Risher, Sarstedt and Ringle, 2019). Both Cronbach's alpha and composite reliability measure the internal consistency of a scale, defined as the measure of how well the items meant to measure a construct on a scale produce similar results. If all items on a scale measure the same construct or idea, the scale has internal consistency or reliability (Cooper and Schindler, 2014).

Cronbach's alpha and composite reliability were therefore used to examine the adequacy of the internal reliability of SET used at NUL. Traditionally, these figures should be above 0.70 for the scale to have internal reliability (Hair et al., 2019, 2011). The results are set out in Table 3.

Table 3. Measures of Internal Reliability of the SET Scale

SET dimension	Cronbach's alpha	Composite reliability
Course delivery	0.867	0.902
Assessment and support	0.853	0.886
Class attendance	0.667	0.821

With the exception of the class attendance dimension of SET (which had mixed results), the internal reliability of all dimensions was above the required figure of 0.70. This suggests that the SET scale used at NUL is reasonably reliable. Some researchers prefer the inter-class reliability over Cronbach's alpha (Morley, 2014), mainly because in SET, multiple evaluators (students) often rate one person (instructor). The inter-class reliability was also acceptable at 0.86.

Reliability is a necessary but not a sufficient characteristic of good scales (Cooper and Schindler, 2014). The following section examines another important characteristic of effective scales – validity.

4.4 Validity of the SET Scale

Validity generally refers to the extent to which the scale measures what it purports to measure (Cooper and Schindler, 2014). There are many forms of validity. For the purpose of this study three popular forms, construct validity, criterion validity, and content validity, were examined.

4.5 Construct Validity of the SET Scale

Construct validity measures the degree to which the theoretical construct assesses what it purports to assess (Cooper and Schindler, 2014, Hair et al., 2010). Convergent validity and discriminant validity are often assessed to evaluate construct validity. Put differently, convergent validity and discriminant validity are components of construct validity.

Convergent validity measures the degree to which the measures of a construct are related (Hair et al., 2010). For instance, a latent construct such as student satisfaction is measured by many observable items (measures) which should be related.

There are several ways in which convergent validity can be confirmed. First, all items should have statistically significant loadings on their latent construct. Second, the average variance extracted should be 0.50 or higher; and finally, the composite reliability should be 0.70 or higher (Hair et al., 2011). Average variance extracted (AVE) measures the level of variance captured by a construct in relation to the level due to measurement error. For instance, AVE of 0.50 means that the items in a latent construct explain 50% of variance in the latent construct; this is considered to be an acceptable figure (Hair et al., 2011).

The standardised loading of items on the SET scale used at NUL are illustrated in Figure 1, and the composite reliability and AVE figures are shown in Tables 3 and 4, respectively.

Although a few items had loadings below 0.70 (Figure 1), all the loadings were significant (t-value of 1.96 or more). Furthermore, the composite reliability of all dimensions was greater than 0.70. Finally, with the exception of factor 2 (student assessment and support), all dimensions had an AVE of 0.50 or more.

Discriminant validity measures the extent to which the measures of different constructs are not related (Campbell and Fiske, 1959; Hair et al., 2010). The study used the Fornell-Lacker criterion to assess discriminant validity. According to this criterion, the AVE of each variable should be greater than the shared variance (the squared correlation) between variables (Hair et al., 2019). The results are shown in Table 4.

Table 4. Average Variance Extracted (AVE) and Shared Variance (SV)

Variable	Attendance	Delivery	Assessment
Attendance	**(0.606)**		
Delivery	0.246	**(0.613)**	
Assessment	0.245	0.417	**(0.493)**

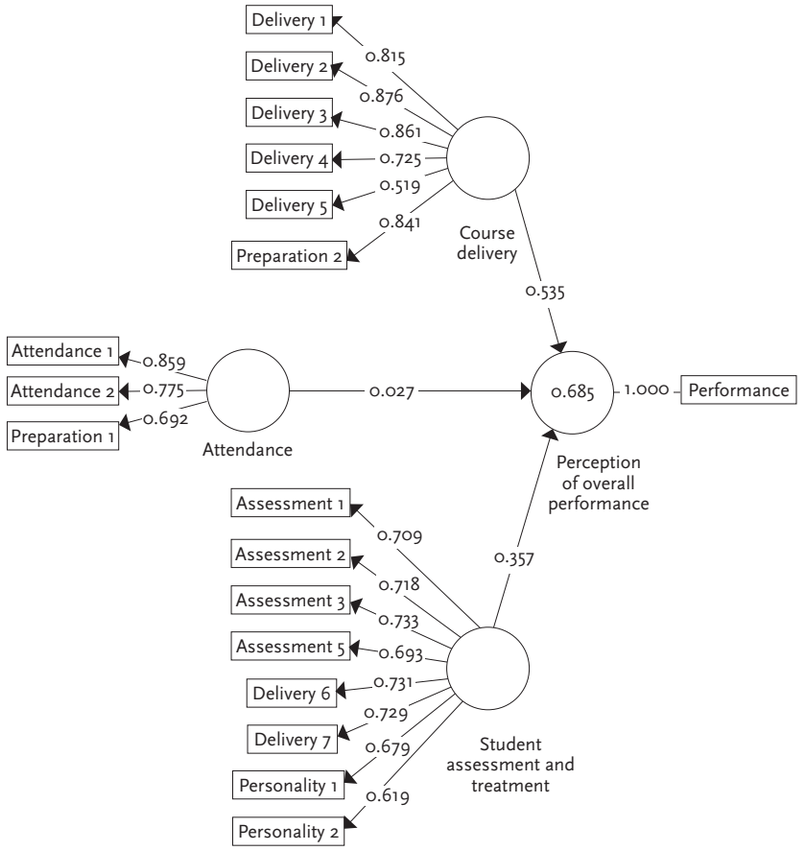
Note: AVE figures are shown in brackets, and other figures represent shared variance (SV) between variables.

As shown in Table 4, the AVE of each dimension was greater than the shared variance between dimensions. Even though the AVE of the 'assessment' dimension was only slightly higher than the shared variance between this dimension and course delivery, the results suggest reasonable discriminant validity between the dimensions of SET used at NUL. Put another way, students' perceptions of lecturer attendance, delivery of lectures and assessment of courses as measured in the SET scale used at NUL could be differentiated from one another.

4.6 Predictive Validity of the SET Scale

Criterion validity, and more specifically, predictive validity, refers to the extent to which the scores of the scale are associated with another variable (Spooren et al., 2013). Traditionally, SET scores have often correlated to students' learning and satisfaction (Clayson, 2009; Spooren et al., 2013; Uttl et al., 2017). However, because the SET scale used at NUL does not assess student achievement, the global indicator of lecturer performance (as assessed by students) was used as a dependent variable that can be predicted by SET scores. On a scale ranging from 1 (not satisfactory at all) to 5 (very satisfactory), students were requested to 'rate the overall performance of this lecturer'. The relationships between the SET dimensions and perceived overall performance of a lecturer are shown in Figure 1.

Figure 1. The Relationships Between SET Dimensions and a Lecturer's Overall Performance



The figure shows that, while effectiveness in course delivery (β = 0.54) and course assessment and support (β = 0.36) significantly predicted performance, class attendance did not. The results suggest that punctuality and 'showing up' are not sufficient to influence students' overall rating of a lecturer. Overall, the three dimensions explained about 69% of variance in perceived performance. This is in line with Spooren et al.'s (2013) conclusion that, "SET research reveals moderate to large positive correlations between SET scores and other indicators of teaching quality".

4.7 Content Validity

The nature of the SET items of the instrument under review suggests that there may be issues with regard to content-related validity (face validity, item validity or sampling validity, i.e., the extent to which the items of an instrument represent the content of the domain being measured) (Spooren et al., 2013, p. 601). As noted by several scholars, there is variation in the SET instruments used, most of which are developed without a clear theory of effective teaching (Clayson, 2009; Hornstein, 2017; Spooren et al., 2013). Thus, although it could be expected that characteristics of effective teachers such as subject knowledge, course organisation, helpfulness, enthusiasm, feedback and interaction with students are now known, "existing SET instruments vary widely in the dimensions they capture" (Spooren et al., 2013, p. 603). While the items included in the SET scale used at NUL cover many items and dimensions of effective teaching, they do not cover items relating to the design and planning of teaching, level of students' learning, course content, and the integration of scholarship, research and professional activities in teaching, to name but a few popular aspects of quality teaching.

In summary, the SET instrument used at NUL may be limited in terms of the extent to which the items represent the content of the effective teaching domain. This is problematic because SET researchers generally agree that the instrument should capture all aspects (dimensions) of good teaching (Spooren et al., 2013, p. 603).

4.8 Gender Bias

The t-differences statistic was used to examine if students evaluated male lecturers more favourably than females. The results are shown in Table 5.

Table 5. Gender and Students' Evaluation

Dimension	Means		t-value	Significance
	Male	Female		
Course delivery	3.81	4.38	8.79	0.000
Lecturer attendance	4.41	4.16	-4.59	0.000
Assessment	3.83	4.00	2.04	0.042
Lecturer performance	3.60	4.09	6.87	0.000

The table illustrates that, on average, female lecturers were judged more favourably than males in three of four dimensions of the SET instrument. Females (mean = 4.38) were rated higher than males (mean = 3.81) in perceptions of course delivery, $t = 8.79$, $p \leq 0.01$. Females (mean = 4.00) were also perceived to be more just in assessment than males (mean = 3.83), $t = 2.04$, $p \leq 0.05$. In terms of performance, female lecturers (mean = 4.09) were again rated higher than male lecturers (mean = 3.60), $t = 6.87$, $p \leq 0.01$. Male lecturers (mean = 4.41) were only rated higher than female lecturers (mean = 4.16) in class attendance, $t = -4.59$, $p \leq 0.01$.

5. Conclusions and Recommendations

The SET instrument is one of the most debated scales in education research. Concerns have been raised about its reliability, validity and inherent biases for the purposes of making personnel decisions in HEIs (Clayson, 2009; Boring et al., 2016; Linse, 2017; Spooren et al., 2013). Furthermore, there are multiple SET instruments, some of which do not derive from theoretical foundations, and are not duly validated (Spooren et al., 2013). Despite these concerns, SET is likely to remain an important instrument to evaluate teaching quality in HEIs (Hornstein, 2017; Linse, 2017).

This article assessed the validity and reliability of the SET instrument used at NUL. While the results suggest that it is reasonably reliable, mixed results were obtained with regard to validity. More specifically, whereas the scale is multi-dimensional with acceptable reliability, convergent validity and predictive validity, there are a few issues with regard to discriminant validity, and serious issues in relation to content validity (face validity, item validity and sampling validity). The gender bias revealed by this study differs from that revealed in the literature (e.g., MacNell et al., 2015; Mitchell and Martin, 2018) as female lecturers were judged more favourably than their male counterparts. This could be attributed to contextual factors.

While the positive attributes may outweigh the negative ones, it would not be prudent to use the instrument in its current form for making serious personnel decisions such as promotions and salary increments. Further recommendations are offered in the final paragraph.

The study suffered from some limitations. Firstly, it was based on a limited number of students registered for one programme in one institution. This limits the generalisability of the results. While the results are in line with those in the literature, future studies could benefit from a larger sample in more institutions. Future studies could also investigate why female lecturers were judged more favourably than their male counterparts. The second limitation is that the study was based on evaluation of the existing scale. This suggests that the inherent weaknesses of the scale are unavoidably also the weaknesses of the study.

A number of recommendations arise from the study's findings. Firstly, SET should be used for formative assessment. As suggested by Penny (2003), it is undeniably an important tool for purposes of improving teaching. Secondly, if SET is used in making decisions in relation to personnel, it should not be employed as the only indicator of teaching quality. Other indicators include observation reports, evaluation by peers, evaluation of education experts, etc. (Little et al., 2009; Spooren et al., 2013). Thirdly, researchers should agree on the definition of teaching quality, and on the instrument that measures it appropriately. Such instrument should be based on theories of learning and teaching, and be validated in the context where it is employed (Spooren et al., 2013). More specifically, the SET instrument used at NUL should be revised, validated, and mainly used for formative purposes. This would increase the applicability and acceptance of this instrument in the institution and similar institutions at the same stage of development.

References

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quinonez, H., and Young, S. L. (2018). Best Practices for Developing and Validating Scales for Health, Social, and Behavioral Research: A Primer. *Frontiers in Public Health 6,* 149.

Boring, A., Ottoboni, K., and Stark, P. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research*. https://www.scienceopen.com (accessed 8 April, 2019).

Cagri, T.M. (2017). Student evaluation of teaching effectiveness in Higher Education. *International Journal of Academic Research in Business and Social Sciences* 7(10), 57-61.

Campbell, D. T., and Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin 56*(2), 56-81.

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*(1), 16-30.

Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research* 51(3), 281-309.

Cooper, D. R., and Schindler, P. S. (2014). *Business research methods* (12th ed.). New York, NY: McGraw-Hill.

Fornell, C., and Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research 18,* 39-50.

George, D., and Mallery, M. (2010). *Using SPSS for windows step by step: A simple guide and reference*. Boston, MA: Addison-Wesly

Hair Jr., J.F., Black, W.C., Babin, B.J. and Anderson, R.E. (2010). *Multivariate Data Analysis: A Global Perspective* (7th Edition). Pearson Education, Upper Saddle River

Hair, J. F., Ringle, C. M., and Sarstedt, M. (2011). PLS-SEM: Indeed a silver bullet. *Journal of Marketing theory and Practice 19*(2), 139-152.

Hair, J. F., Risher, J. J., Sarstedt, M., and Ringle, C. M. (2019). When to use and how to report the results of PLS-SEM. *European Business Review 31*(1), 2-24.

Higher Education Act 2004, downloaded from http://www.che.ac.ls/wp-content/uploads/2019/01/Higher-Education-Act-2004.pdf on 8 April 2020

Hornstein, H. A. (2017). Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance. *Cogent Education 4*(1), 1-8.

Linse, A. R. (2017). Interpreting and using student ratings data: Guidance for faculty serving as administrators and on evaluation committees. *Studies in Educational Evaluation 54*, 94-106.

Little, O., Goe, L., and Bell, C. (2009). *A Practical Guide to Evaluating Teacher Effectiveness*. National Comprehensive Center for Teacher Quality.

MacNell, L., Driscoll, A., and Hunt, A. N. (2015). What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education 40*(4), 291-303.

Marks, R.B. (2000). Determinants of student evaluations of global measures of instructor and course value, *Journal of Marketing Education 22*(2), 108-11.

McBean, E. A., and Al-Nassri, S. (1982). Questionnaire design for student measurement of teaching effectiveness. *Higher Education 11*(3), 273-288.

Mitchell, K. M., and Martin, J. (2018). Gender bias in student evaluations. *PS: Political Science and Politics 51*(3), 648-652.

Morley, D. (2014). Assessing the reliability of student evaluations of teaching: choosing the right coefficient. *Assessment and Evaluation in Higher Education 39*(2), 127-139.

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.

Siebert, J. and Kunz, R. (2016). Developing and validating the multidimensional proactive decision-making scale. *European Journal of Operational Research, 249*, 864-877.

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

Uttl, B., White, C. A., and 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.