Iranian Journal of Comparative Education, 2026, 9(1), 4056-4076

Research Article **DOI: 10.22034/IJCE.2025.528182.1660** http://journal.cesir.ir

Examining the Optimal Length of SET Forms: Lessons from the Iranian Higher Education

Ali Mohammad Fallahi Barzoki [1]
Esmaeel Ali Salimi [2](Corresponding author)
Kamran Sheivandi Choliche [3]

ARTICLE INFO

Received: 03 June 2025
Revised: 22 July 2025
Accepted: 04 November 2025
Online: 21 April 2026

K E Y W O R D S

Evaluation System
Faculty Appraisal
Higher Education
SET
Teaching Evaluation

ABSTRACT

Universities and institutions of higher education increasingly rely on student evaluations of teaching (SET) to enhance both institutional and educational quality. SETs are commonly implemented worldwide using standardized survey instruments. These instruments are meticulously constructed to align with validated indicators of instructional effectiveness and the professional competencies integral to the teaching vocation. Comparative analyses of evaluation instruments across leading universities indicate significant convergence in assessment criteria, yet the optimal length of SET forms remains underexplored. Analyzing approximately 27,000 student surveys at the tertiary level, this study identifies the ideal number of SET items through exploratory and confirmatory factor analyses, leading to the design of five short-form scales for evaluation of teaching. Statistical comparisons, including linear regression and means tests, demonstrate that two of these concise forms yield estimates of instructors' overall mean scores comparable to those derived from full-length evaluations. These findings suggest that streamlined SET instruments—consisting of four to five rigorously-validated items—can maintain methodological integrity while enhancing student attentiveness as well as improving psychometric properties, thereby ensuring greater reliability and validity in faculty evaluations.

[1] PhD Student, Department of English Language & Literature, Faculty of Persian Literature and Foreign Languages, Allameh Tabataba'i University, Iran, Email: fallahibarzoki_ali@atu.ac.ir
[2] Associate Professor, Department of English Language & Literature, Faculty of Persian Literature and Foreign Languages, Allameh Tabataba'i University, Tehran, Iran, Email: easalimi@atu.ac.ir
[3] Associate Professor, Department of Educational Psychology, Faculty of Psychology and Educational Science, Allameh Tabataba'i University, Iran, Email: sheivandi@atu.ac.ir

1. Introduction

Organizational elements enhance lecture effectiveness in educational settings (Hanh, 2023) and student evaluation of teaching (SET) as one of these factors will significantly influence the teaching process. Within the Iranian higher education system, SET results play a critical role in various personnel decisions, including contract extensions, employment status, and academic rank promotions. The administrative procedure involves aggregating and analyzing SET mean scores over multiple semesters, with interpretations guided by institutional benchmarks such as departmental and faculty-wide averages. This systematic approach ensures that evaluative outcomes contribute meaningfully to academic personnel assessments while maintaining consistency across institutional standards.

However, numerous uncertainties emerge concerning students' evaluation of their professors' teaching performance (Gelber, 2020). Such evaluations are invariably susceptible to multifaceted and systemic forms of bias, thereby leading a number of educational theorists and assessment specialists to advocate, with increasing conviction, for a shift in their conceptualization, utilization, and methodological integrity (Hudson, 2025).

These concerns pertain to students' capacity to credibly comment on specific instructional attributes, such as a professor's expertise in the subject matter; the permissibility for universities to mandate student participation in filling SET forms; whether it is academically appropriate to schedule and conduct the SET process prior to the end-of-semester exams; preserving the anonymity of students during this evaluation; and giving weight to the results of SET forms when making decisions about faculty contract renewals, employment status changes, or promotions, etc. Among these issues, the length of each question or item in the questionnaire or the scale of response to each question or item (i.e., 5-point versus 7-point Likert scale) affects participants' responses; hence the credibility of SET forms (Dillman, 2020; Willis, 2020). Obviously, such variables, even the time allocation to answer SET questions, can exert significant impacts on the final grades assigned to instructors. For instance, in some universities, the survey is restricted to a single week preceding final examinations. This limited timeframe may hinder students' ability to recall instructional activities accurately, thereby compromising the reliability of their assessments. Furthermore, when students are required to complete multiple evaluations in a condensed period—often within a single sitting—cognitive fatigue may lead them to rely on generalized impressions rather than systematically appraising pedagogical effectiveness based on established criteria.

Reflecting on these technical aspects vis-à-vis SET forms is of high importance because these details cannot systematically affect students' assessments of their professors. A significant hypothesis in this context proposes that short SET forms enable students to concentrate more effectively on each individual item, thereby enhancing the validity and reliability of their feedback (Gursoy & Umbreit, 2005). This consideration is particularly significant given that, in most universities worldwide, teaching evaluations are now conducted electronically, replacing the traditional in-person, pencil-and-paper method administered within classrooms. The modern approach allows students to participate in evaluations across diverse contexts—such as while watching a movie or commuting on the subway—where external distractions may diminish their attentiveness. Consequently, subjecting students to an extensive set of evaluation items in such conditions risks inducing both conscious and unconscious inattentiveness, thereby compromising the integrity and reliability of the assessment process.

Nevertheless, merely reducing the number of items on a SET form does not ensure favorable outcomes. This limitation arises from the influence of various other factors on the validity and reliability of such forms, including the number of courses a student enrolls in during a semester. Specifically, given that most universities now implement electronic teacher evaluations, it is a moot point whether students complete evaluation forms for all instructors consecutively on the same day and hour or if they allocate specific times for each evaluation. There has been a paucity of independent research in this domain; however, it appears that students, particularly within the context of Iranian higher education, often opt to complete all evaluations simultaneously.

2. Literature Review

SETs have become an established instrument in higher education for assessing instructional effectiveness among faculty members (Gilbert & Gilbert, 2025). Within institutions of higher education, professorial performance is subject to review, and SETs invariably constitute a significant component of these assessment protocols (Gelber, 2020). Such student-teacher evaluations have been a subject of sustained scholarly inquiry among academic administrators and pedagogical experts for over five decades. Scholars broach many subjects and raise many questions in regard to SET forms and the factors which wield influence on them. Student-instructor interaction, non-anonymity repercussions, gender bias, program evaluation, psychological interventions, etc. (Boswell, 2024; Amini et al., 2019; Khanjani et al., 2017; Sahragard & Farangi, 2017) as well as all those posed in the introduction. Although a whole host of research (Lakeman et al., 2021; Landrum & Braitman, 2008; Park & Cho, 2022) provide us with a gamut of answers, we

still do not know exactly how students complete the SET forms. We are not cognizant of the extent to which they allocate sufficient time to evaluate each item, engage in thoughtful consideration of the text associated with each item, try to thoroughly review the professor's performance during the semester before grading, evaluate all courses in a few minutes at a time, or carefully complete each one at a specific time. One of the most interesting and thought-provoking findings for university administrators can be the research conducted by Spooren & Christiaens (2017), which showed that students who believe the teaching evaluation system in the institutions is efficient and valuable give higher scores to the professors. This indicates, regardless of all ambiguities, the importance of university authorities' efforts to continuously improve the SET process.

The administrative procedure of universities requires that SET forms be usually completed electronically about two weeks before the commencement of the end-of-semester exams. Normally, students undergo considerable stress and experience a state of free-floating anxiety caused by exam-periods which may in turn exert deleterious effects on the SET process. Consequently, this will lead to whether students should evaluate all their courses at once or they had better allocate separate time for each course assessment. In rare cases, it has been observed that some students even fill in the SET forms for each other; or some others protest why they have to waste their time choosing from among some options where they have to read these options one by one! Instead, they would rather click an icon simply, with minimal energy and time! Such examples raise the question of how to make a more accurate evaluation of professors' teaching. Obviously, one of the primary considerations is the length of the SET forms. Logically, it can be assumed that if students know they need to spend only few minutes to complete a few items for each course, they will probably go through this process more accurately.

In practice, SET forms wielded by various universities differ enormously. For example, SET forms used in 15 Iranian universities include a range of 9-20 items. This is of immense significance because each student in Iran's higher educational system usually takes about 6 courses in each semester at the BA levels, and if they are supposed to complete 20 items for each course, the number of items will be copious. Similarly, in postgraduate programs, students are typically required to complete approximately twelve courses; akin to undergraduate curricula, faculty members are mandated to conduct electronic evaluations of each course upon conclusion of the academic term. This issue becomes critical when students complete all the items just in a few minutes. In other words, students can enroll in a course load ranging from 6 to 12 courses each semester; however, this is contingent upon their GPA. These students are obligated to undertake an electronic evaluation form of the instructor at the conclusion of each semester. Given the

comprehensive nature of such forms employed for each course, active student engagement in filling out SET forms necessitates a considerable investment of both time and cognitive resources.

In fact, what exactly is measured in SET forms? A plethora of findings highlight the necessity of rethinking and restructuring of such forms. For example, Uttl, White and Gonzalez (2017) found that SET results had nothing to do with the amount of actual learning in the course. Elsewhere, Park and Cho (2022) concluded that students will punish or praise their instructors depending on whether they are assigned their ideal grade or not.

Uttl (2024) emphasizes that SET instruments are not valid indicators of instructors' teaching effectiveness. He argues that these evaluations fail to accurately assess teaching quality, and students do not achieve greater learning outcomes from professors who receive higher ratings. On the other hand, Graf (2024) contends that SET scores serve as valid indicators of student satisfaction with both courses and instructors, making their application in career decision-making both reasonable and justifiable.

Similarly, many university professors have corroborated such findings or beliefs based on their own personal experiences; and therefore, some are against the entire SET system and consider it a factor in the decline of academic standards. Of course, this does not necessarily mean that SET process lacks validity or is of low value. On the contrary, researchers can find almost a cornucopia of invaluable data vis-à-vis teaching (Buller, 2012). However, authorities are to provide the most efficient tools to carry out such evaluation processes. It seems that one of the most important and economical tools is a short SET form which focuses on primary teaching skills of professors so that students can objectively complete it in the shortest possible time. The Classical Test Theory (CTT) underlying assumptions advocate for the creation of longer tests as opposed to shorter ones (Cohen, Schneider & Tobin, 2022).

Conversely, results of some studies (Burisch, 1998; Shrout & Yager, 1988) demonstrated that the sensitivity and specificity of the scale remained stable with a decrease in length of the SET forms. Liebe et al. (2015) also detected that the SET lengths which endure longer times for completion can lead to a more negative evaluation. In addition to the length of the forms, many other factors such as readability, vocabulary difficulty, term ambiguity, grammatical complexity, etc. can impose a respondent burden, hence poor data quality (Dykema et al., 2020).

Given the SET literature, there appears to be a paucity of independent studies focused on determining the optimal length of evaluation forms and therefore the current research aims to test the hypothesis whether short SET forms can measure the effectiveness of professors' teaching as accurately as extended forms.

All in all, this study addresses the critical inquiry regarding the ideal quantity of items in the SET form that higher education administrators ought to contemplate to ensure that students provide the most accurate assessments of their professors. The findings from this research will contribute to the establishment of more refined operational protocols for the faculty evaluation system.

3. Research Method

This research is based on correlation and structural equation modeling, as it employs a series of exploratory and confirmatory factor analyses to identify the optimal length of SET forms. This methodological approach has been employed to standardize measurement scales in research conducted by Garrido et al. (2024) and Huang & Liu (2024). Overall, the research methodology employed in this study is primarily correlational in nature, wherein the data have been analyzed using exploratory factor analysis, confirmatory factor analysis (structural equation modeling), and linear regression.

The statistical population of the study consists of 26500 SET forms administered during the academic year 2017 at the University of Kashan, with the data extracted from the database of the educational platform of the university (written consent letter no. 1248300). This university was selected due to its meticulous and electronic implementation of a SET system over recent years, where the SET form has undergone regular reviews. One of the administrators' primary concerns has been the creation of suitable and effective forms at this university.

Although, the larger the sample size is, the more consistent the estimates will be (Thompson, 2004), the psychometric foundations and the validations of the scales in this research, in addition to analyzing the entire data set, led to samples of 5, 10, and 15 percent of the dataset which were randomly selected in three stages using software, where all analyses were carried out and reported separately.

The selection of small samples was based on the understanding that, despite the stipulations of Iranian higher education laws and regulations, which dictate that professors' evaluation scores are typically assessed over a span of 8 to 10 academic semesters, the frequency of these evaluations is significantly lower than the overall size of the statistical population in this study. Consequently, to enhance the precision of the analyses, the researcher incorporated both the complete statistical population and smaller samples, thereby allowing for a more realistic and practical simulation of the conditions.

The data collection tool in this study was the official institutional form for assessing the quality of instructors' teaching. The items of this form are presented in Table 1.

Table1. Items in SET Forms during the Specified Semesters

	SET Form Items
1.	Proficiency in articulating, conveying, and disseminating course materials
2.	Provision of pertinent examples and appropriate exercises to enhance understanding
3.	Instructor's flair for the subject matter and course content
4.	Instructor's capacity to address academic inquiries and respond effectively to questions
5.	Presentation of a comprehensive course outline at the commencement of the semester, while adhering to it throughout the term
6.	Maintenance of coherence and continuity in the delivery of course content
7.	Generating motivation and encouraging affective engagement among students vis-à-vis study and research endeavors
8.	Implementation of ongoing and relevant assessments aligned with course objectives
9.	Consistent attendance of instructor in accordance with the established class schedule
10.	Classroom management method and optimal utilization of instructional time
11.	Instructor's office-time management so as to be available for consultation outside of scheduled class hours
12.	Professional demeanor and promotion of mutual respect within the learning environment
13.	Adherence to ethical standards and sensitivity to cultural norms within the academic context
14.	Commitment to ethical principles and consideration of values in instructional practices
15.	Instructor's social behavior to garner mutual respect
16.	Promotion of self-efficacy and enhancement of student dignity and identity within the classroom
17.	Overall assessment of instructor

In most Iranian universities, the teaching assessment process is conducted approximately one week prior to the final examinations. During this period, students individually complete electronic assessment forms for each of their courses. This 9-point scale allows students to assign scores ranging from 12 to 20 for each item pertaining to the course instructor. The objectives, scale, and methodology of the data analysis ensured that the identities of both professors and students were kept confidential during all phases of data screening and statistical evaluation, thereby eliminating the necessity to concentrate on any particular individual.

4. Findings

Exploratory factor analysis (EFA) was initially conducted on the data. Kogar (2020) identifies factor analysis as one of the methods for developing concise forms whereas Dimitrov (2011) introduces EFA as an instrument wielded when there is not ample "theoretical or empirical information to hypothesize how many constructs (factors) underlie the initial set of items and which items form which factor" (p. 45).

Although effective teaching patterns or manners have already been clearly defined and acknowledged (Adams & Engelmann, 1996; Borich, 2017; Burden & Byrd, 2019; Kahn & Anderson, 2019; Kauchak & Eggen, 2017; King, 2022; Slavin, 2018), universities might take into serious consideration their instructors' various aspects of professional and institutional performance. Therefore, in this study EFA was prioritized to identify components based on the internal correlations among the items rather than categorizing items based on theoretical foundations of teaching. EFA was used and five components were identified from the items. Components 1-3 consist of four items, while components 4 and 5 are derived from six. The principal components analysis was employed again for item aggregation, whose results for each component have been reported alongside their reliability coefficients as measured by Cronbach's alpha.

Table 2. EFA Results on SET Items

Samples	**N**	**Forms**	**Items**	**CR**	**EFA***		**CFA***				
					ESSL	KMO	TLI	CFI	GFI	AGFI	RMR
Population	27351	Extended	17	0.981	77.202	0.98	0.286	0.381	0.901	0.871	0.19
		Short Form	4	0.94	84.806	0.853	0.884	0.961	0.991	0.954	0.037
		Short Form	4	0.951	87.166	0.857	0.685	0.859	0.973	0.866	0.08
		Short Form	4	0.936	83.897	0.867	0.997	0.999	1	0.999	0.008
		Short Form	6	0.962	84.194	0.935	0.751	0.85	0.973	0.936	0.08
		Short Form	6	0.949	79.723	0.904	0.464	0.679	0.927	0.829	0.245
1	1355	Extended	17	0.99	75.834	0.976	0.268	0.36	0.895	0.866	0.186
		Short Form	4	0.931	82.911	0.842	0.904	0.968	0.992	0.961	0.034
		Short Form	4	0.953	87.849	0.856	0.661	0.887	0.971	0.853	0.084
		Short Form	4	0.933	83.266	0.864	0.996	0.999	0.999	0.997	0.011
		Short Form	6	0.962	84.13	0.932	0.749	0.85	0.972	0.934	0.08
		Short Form	6	0.942	77.567	0.892	0.438	0.663	0.921	0.815	0.243
2	2694	Extended	17	0.982	77.783	0.979	0.271	0.362	0.895	0.865	0.168
		Short Form	4	0.94	84.804	0.855	0.88	0.96	0.99	0.951	0.042
		Short Form	4	0.949	86.87	0.851	0.614	0.871	0.966	0.83	0.097
		Short Form	4	0.941	85.016	0.869	1.001	1	1	0.999	0.007
		Short Form	6	0.96	83.553	0.932	0.711	0.827	0.966	0.922	0.092
		Short Form	6	0.952	80.885	0.91	0.489	0.693	0.931	0.838	0.204
3	4093	Extended	17	0.892	77.724	0.979	0.268	0.36	0.895	0.866	0.186
		Short Form	4	0.941	84.972	0.855	0.904	0.968	0.992	0.961	0.034
		Short Form	4	0.951	87.398	0.856	0.661	0.887	0.971	0.853	0.084

Short Form	4	0.938	84.347	0.867	0.996	0.999	0.999	0.997	0.011
Short Form	6	0.964	84.758	0.936	0.749	0.85	0.972	0.934	0.08
Short Form	6	0.949	79.951	0.902	0.438	0.663	0.921	0.815	0.243

Note: ESSL is Extraction Sum of Squared Loading; and CR is Composite Reliability.

*All EFA and CFA statistics were significant at the 99% confidence level.

The Kaiser–Meyer–Olkin (KMO) measure verified the sampling adequacy (KMO values for all models were > 0.85), which is well above the acceptable limit of 0.5. Bartlett's test of sphericity was also significant for all models, indicating that correlations between items were sufficiently large for Principal Components Analysis (PCA). Table 2 further shows the extraction sum of squared loadings, i.e., "the amount of variance in the items that can be explained by that particular principal component", (Pett et al., 2003, p.92).

A noteworthy finding is that the variance explained by the entire scale (encompassing all 17 items) in all samples is lower than that of the other five models and does not exceed 0.77. The plausible reason for this is that the comprehensive scale measures diverse facets of instructor's performance, which do not necessarily pertain to educational dimensions. Naturally, items exhibiting higher internal correlation possess greater validity, hence superior measurement capability for the latent variable under consideration. In accordance with the data presented in Table 2 and considering the crucial parameter of the number of the items, the concise SET form #3, comprising four items, is identified as the most appropriate substitute for the complete scale.

One of the most popular estimates of internal consistency is Cronbach's α. This index has also been reported separately for each of the component combinations. The alpha coefficient is influenced by the scale length or the number of items (Kogar, 2020); thus, it is naturally observed that, in all instances, the alpha value for the "total scale" is higher. Overall, if $\alpha \geq 0.9$, the internal consistency is considered to be excellent, and so is observed for all models in the study.

Generalized Least-Squares (GLS) estimation was used for CFA. For large samples, GLS provides the same goodness of fit as ML (Brown, 2015; Mulaik, 2009). It is important to highlight, as discussed in the methodology section, that both model x2 and modification indices exhibit sensitivity to sample size (Brown, 2015). Thus, conducting replication studies with various samples would provide evidence for the stability of the findings (Schreiber et al., 2006). Therefore, all analyses in the present study were conducted on three random samples from the population.

Several model fit indices were used to examine the goodness-of-fit of the model with the given dataset, including goodness-of-fit index (GFI), adjusted goodness-of fit index (AGFI), Tucker-Lewis Index (TLI), and comparative fit index (CFI).

However, some indices were not reported in Table 2. As an instance, since CMIN test must be used with some care, as it has been shown to be both sensitive to sample size and not robust to departures from multivariate normality by the data. Or, one of the common indices is RMSEA, which is sensitive to low degrees of freedom (Finch, Immekus & French, 2016).

In forms #1 to 5, the number of items was either 4 or 6, resulting in low degrees of freedom; thus, RMSEA was not reported.

Factor analysis is primarily concerned with a particular aspect of construct validity evidence, specifically the internal structural element that assesses the degree to which the scoring system corresponds with the construct domain's structure (Messick, 1995). While strong factorial evidence is crucial, it is not sufficient on its own to establish validity (McCoach et al, 2013). The importance of factors must be evaluated in terms of their stability across different samples and methods, in addition to the significance of their relationships with external criteria (Watkins, 2021).

Some experts revealed that the expected value of the GFI index tended to increase with sample size (Mulaik, 2009). Numerous factors must be taken into account when interpreting the indices outlined in specialized literature (Dimitrov, 2011; Marsh & Bella, 1994; Niemand & Mai, 2018; Preston & Colman, 2000). In this context, to enhance the reliability of the findings, the average of the pertinent indices for each short form has been individually computed and reported.

Table 3. EFA/CFA Mean Indices for All Models

	Reliability	**ESSL**	**KMO**	**TLI**	**CFI**	**GFI**	**AGFI**	**RMR**
Total Scale	0.981	77.202	0.980	0.286	0.381	0.901	0.871	0.190
1	0.938	84.373	0.851	0.893	0.964	0.991	0.956	0.036
2	0.951	87.320	0.855	0.655	0.876	0.970	0.850	0.086
3	0.937	84.131	0.866	0.997	0.999	0.999	0.998	0.009
4	0.962	84.158	0.933	0.740	0.844	0.970	0.931	0.083
5	0.948	79.531	0.902	0.4573	0.674	0.925	0.824	0.233

The proximity of the TLI, CFI, GFI, and AGFI indices to 1, along with the RMR's proximity to 0, suggests an enhanced fit and utility of the model. In other words, the values of CFI > .95 and TLI > .95 indicate a good model-data fit (Hu & Bentler, 1999; Xia & Yang, 2019).

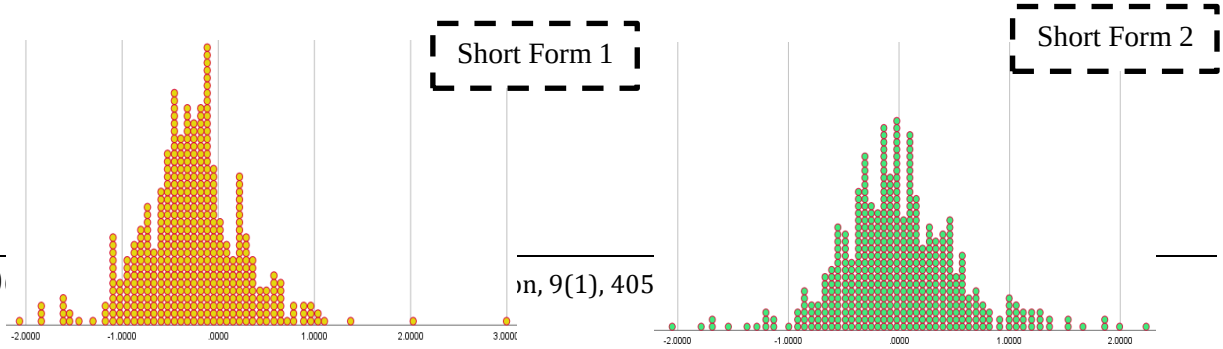
As detailed in Table 3, a comparative analysis reveals that the indices for forms #2 and 3, each comprising four items, are more favorable. However, could these two short forms, with only four items each, exhibit the same measurement power as the full-length form? Calculating the correlation between the short forms and the full-length form demonstrates that all correlations are

significant. Given the consistency in semester conditions, course conditions, class composition, and the teaching evaluation form itself, it is logical that a high evaluation score on the full-length form corresponds to a high score on the short forms as well. Therefore, selecting a more appropriate criterion could enhance the accuracy of the analyses. To address this issue, the evaluation mean score of each instructor was calculated over a 10-semester period. The average number of students evaluating the instructor in a semester was 150, but the average number of evaluations was approximately tenfold, making the 10-semester mean score a relatively valid indicator of an instructor's effectiveness and efficiency. Subsequently, the correlation between the scores obtained from the short forms with a 10-semester evaluation mean was calculated. Given the significance of all correlations, simple linear regression was employed. Here, the question was to determine which of the short forms (previous semester evaluations) could predict the long-term evaluation mean of the instructor. The results are presented in Table 4.

Table 4. Summary of Regression for Short Forms' Scores Predicting10-Semester Evaluation Mean

Model		Unstandardized Coefficients		Standardized Beta	T	Sig.
		B	Std. Error			
Short Scales	(Constant)	6.544	0.488		13.419	0.000
	1	0.098	0.147	.1100	0.668	0.505
	2	0.536	0.268	0.704	2.002	0.046
	3	-0.149	0.166	-0.189	-0.901	0.368
	4	-0.018	0.318	-0.023	-0.056	0.955
	5	0.176	0.257	0.206	0.682	0.496

As shown in Table 4, the only significant standardized coefficient pertains to the short form #2. In fact, the teaching evaluation score of the instructor based on this short form possess the capacity to predict the overall evaluation score of the instructor over ten semesters. In addition, the simple mean of all the instructors' evaluations was calculated. Then, the differences between the short form evaluation scores and the mean of the long form, as well as the ten-semester evaluation mean of the instructors, were calculated. The distribution of score differences is depicted in graphs for each short form.



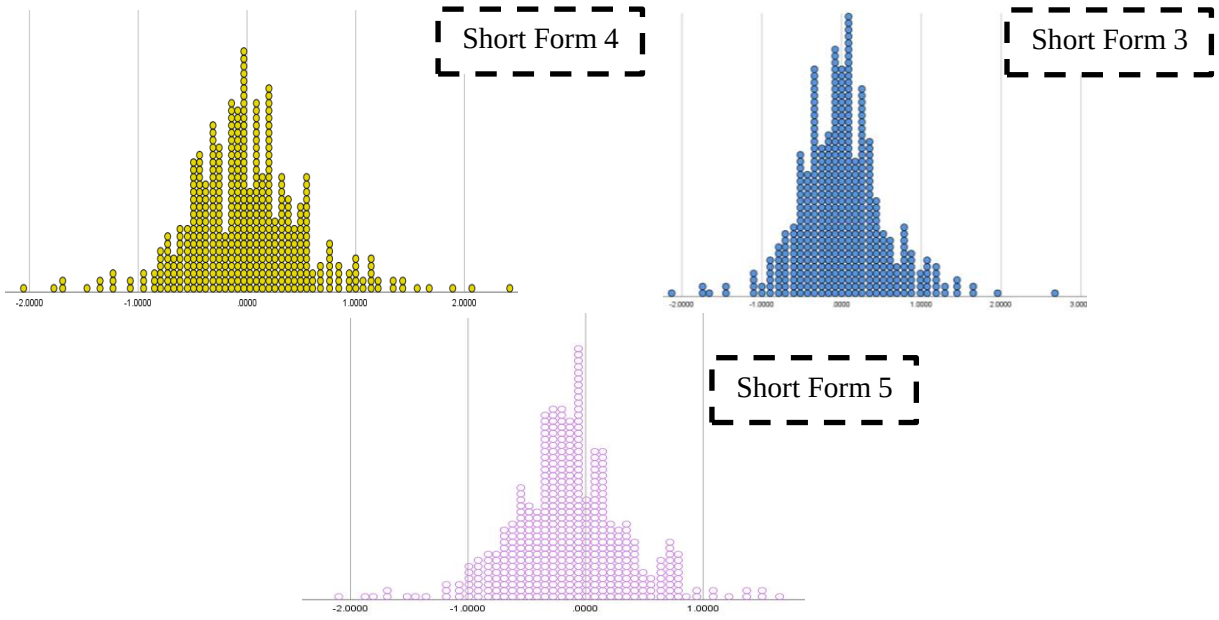


Figure 1. Distribution of the Difference between Short-form Evaluation Scores and the 10-semester Evaluation Mean

The graphs illustrating the 5 short forms reveal that the scores for these forms vary in relation to the instructor's 10-semester evaluation mean, oscillating between +1 and -1. Notably, the outlier data, which appear at the extremes of the chart, primarily relate to part-time instructors or those nearing retirement. This observation prompted researchers to contemplate the separation of data for part-time instructors from that of full-time instructors. Thus, a hypothesis was formulated and tested, positing that there is no significant difference between the evaluation scores of the short forms and the 10-semester evaluation mean. The findings are detailed in Table 5.

Table 5. Results of One-Sample t Test for Comparing the Mean Difference between Short Form Scores and the Desired Test Value (Zero)

Statistics	Teacher Status	N	mean	t	Sig.
Total Scale	Part Time	195	-0.050	-1.439	0.152
	Full Time	247	-0.093	-3.124	0.002
1	Part Time	195	-0.276	-7.134	0.000
	Full Time	247	-0.269	-8.687	0.000
2	Part Time	195	0.033	0.891	0.374
	Full Time	247	0.037	-1.098	0.273
3	Part Time	195	0.028	0.736	0.462
	Full Time	247	-0.008	-0.268	0.788
4	Part Time	195	0.040	1.086	0.279
	Full Time	247	-0.049	-1.499	0.135
5	Part Time	195	-0.135	-3.706	0.000
	Full Time	247	-0.139	-4.622	0.000

The data presented in the preceding Table suggests that a lack of differences is desirable in this context, reflecting the effectiveness of the short scale. As shown in Table 5, the score variations for short forms #2, 3, and 4 are not statistically significant when compared to the target value of zero. The findings from the confirmatory factor analysis (Table 2), regression analysis (Table 4), and mean comparison test (Table 5) lead to the conclusion that the short form #2, followed closely by short form #3, exhibit the most favorable indices, respectively. The items of these short forms are detailed as follows:

Items	Short Scale 2
1	Provision of pertinent examples and appropriate exercises
2	Instructor's flair for the subject matter and course content
3	Generating motivation and encouraging affective engagement among students vis-à-vis study and research endeavors
4	Implementation of ongoing and relevant assessments aligned with course objectives

Items	Short Scale 3
1	Proficiency in articulating, conveying, and disseminating course materials
2	Instructor's capacity to address academic inquiries and respond effectively to questions
3	Maintenance of coherence and continuity in the delivery of course content
4	Classroom management method and optimal utilization of instructional time

Figure 2. Items from Two Short SET Forms

What makes the results of these short forms credible? What do these short forms assess? Can deans of universities confidently implement a SET system based on such concise instruments? The forthcoming section will address and clarify these inquiries.

5. Discussion

The results of SET forms regarding professors' teaching effectiveness are shaped by several essential factors. To guarantee that the feedback obtained from students is both valid and reliable, it is imperative to satisfy various critical conditions, including scale length and timing (Clift, Hall, & Turner, 1989). Additionally, factors such as students' initial interest in the course and its classification as either required or elective play a significant role (Linask & Monks, 2018). The primary concern addressed in this study pertains to a significant operational and executive aspect of SET forms vis-à-vis professors' teaching effectiveness, specifically the length of the SET form as a critical variable that can systematically influence the outcomes of these evaluations.

The findings of the current research indicate that a concise SET form comprising four relevant items can be as effective as a more comprehensive version for assessing instructors through student evaluations. This approach's feasibility and efficacy have been explored in previous studies, including those by Sherout and Yager (1988) and Gonzalez (2021). Furthermore, Fitzpatrick and Yin (2001) have recommended, based on their analyses, that evaluation instruments should ideally include a minimum of eight 6-point items or at least twelve items with four or more scoring points each.

The SET forms utilized in this research consisted of four items; however, it is important to highlight that our evaluation instrument employed a 9-point scale. The extent of the scale is inherently linked to the reliability of the instrument (Kutscher & Eid, 2020; Niemi et al., 1986; Taherdoost, 2019); Scales that encompass a greater number of items tend to produce scores with increased variability (Finch, Immekus & French, 2019). Consequently, alongside the substantial sample size in this study, the extended scale length has further enhanced the reliability of the short forms.

The primary rationale for the effectiveness of the short forms lies in their ability to evaluate the behaviors or attributes associated with effective teaching. For example, the items in short form #1 can be aligned with the principles of direct instruction (Adams & Engelmann, 1996; Kauchak & Eggen, 2017). Over time, the development of effective teaching skills plays a crucial role in enhancing students' learning experiences and improving their academic performance (Borich, 2017; Burden & Byrd, 2019). Essentially, the items in short form #1 assess the core teaching skills of instructors—skills that students consciously or unconsciously recognize as beneficial for their learning, thereby influencing their assessments of the instructors.

The items in Form #1 align with the criteria that specialists have identified for evaluation instruments. For example, the four items in Form #1 correspond to the organizational dimension as

classified by Gursoy & Umbreit (2005), Marks (2000), and Marsh and Roche (1997). Thus, it can be inferred that while a short SET form may not encompass all aspects of teaching or provide professors with comprehensive feedback (Dodeen, 2013), it nonetheless demonstrates significant validity and reliability in assessing the fundamental dimensions of effective teaching.

Psychometric approaches focused on estimating a latent variable, which provides a mechanism for inference and interpretation of the construct. With a latent variable, researchers could estimate precise scores that could be ascribed a valid interpretation (Gonzalez, 2021). In explaining the efficacy of short-form items, such as those utilized in forms #1 and 2 of the current study, it is pertinent to reference the significant meta-analysis conducted by Rolstad, Adler, and Rydan (2011). Their findings emphasize that the content of the form should be prioritized over its length.

It is essential to recognize that the conciseness of SET forms may lead to inadequate feedback for educators regarding their instructional practices. A robust teacher evaluation system is undeniably beneficial when it facilitates the enhancement of teaching quality (Buller, 2012). However, if the evaluation forms are drastically shortened, instructors will not receive comprehensive feedback from students on critical indicators of effective teaching, such as proper use of learning probes, fulfilment of course objectives, etc. Similarly, this item reduction can seriously question and significantly undermine the effectiveness of the evaluation system. Furthermore, there is a concern that a substantial decrease in the number of items on the evaluation form may convey to students that the evaluation process is of little importance, which could have detrimental effects. Nederhand et al. (2022) claim that even when students are provided with a succinct rationale for the significance of the SET forms, their participation and motivation do not immensely improve. Thus, any decision to streamline SET forms should be accompanied by strategies aimed at increasing both the quantity and quality of student engagement in the evaluation process. Cunningham and MacGregor (2006) assert that adherence to conventional evaluation metrics for pedagogical adjustments may cause students to underestimate values for some factors, ultimately resulting in no enhancement of overall teaching evaluations. In the conclusion section, some strategies for addressing this issue are outlined.

6. Conclusion

What are the implications of the current study for university authorities and the development of instructor evaluation systems? Given the shared foundational logic and prevalent assumptions underpinning SET systems across diverse academic institutions, coupled with the notable

similarities in evaluation instruments (survey forms, in particular) and their implementation methodologies, the findings of this research are expected to warrant careful consideration and present pragmatic implications for academic administrators. This study was conducted to support university administrators in optimizing the structure and length of SET forms, with the dual aim of enhancing response precision and attenuating the prevalence of systematic biases empirically associated with student evaluations of teaching (Heffernan, 2021; Rolph, King & Larde, 2023).

The main rationale for our recommendation is that the quantitative analyses employed by numerous university systems concerning student evaluation data tend to be rather superficial and overly simplistic. For example, some systems focus solely on the SET mean scores from all students who have assessed the instructor, while some others do make a distinction between undergraduate and graduate courses; however, the predominant emphasis remains on the SET mean scores of the instructor. At best, this mean is juxtaposed and compared with mean scores of the respective department or faculty.

In contrast, Buller (2012) asserts that quantitative analysis alone cannot provide reviewers with a complete understanding of the situation, as numerical data often lacks the necessary context.

The extended length of SET forms appears to provide little additional insight into the teaching conditions and circumstances faced by instructors. In effect, university administrators ought to consider implementing measures aimed at shortening these forms. Several strategies are proposed herein to address this issue.

The initial strategy involves enhancing the accuracy and precision of student responses by categorizing the questions/items into distinct categories. In fact, to address the concerns of researchers such as Dodeen (2013), who emphasize the importance of the multidimensionality of evaluation forms, one alternative is to adopt a temporal approach to multidimensionality. This would involve implementing various forms over time, rather than relying on a single form that encompasses multiple dimensions.

Upon logging into their portal to fill out the SET forms, students would encounter one of these categories at random. This approach is particularly suitable for assessing courses with a substantial enrollment, such as those exceeding 30 students.

When examining the outcomes of multiple semesters of SET surveys in relation to administrative decisions, such as academic promotion, the enrollment figures for each course become largely inconsequential. This is due to the fact that the professor has been assessed by a substantial number of students across various dimensions over an extended timeframe.

The second strategy focuses on distinguishing between aspects of the instructor's professional conduct and those related to effective teaching practices. Numerous SET instruments incorporate items that address the instructor's avoidance of biases related to gender and religion, as well as their respect for students. However, these items often do not align well with established measurement principles. Evaluating such attributes necessitates a separate assessment tool, comprising multiple items because abstract concepts such as "respect for students," "adherence to values," and "ethics" can be interpreted differently across diverse student populations. In practice, items containing abstract value concepts tend to lack diagnostic efficacy, as each holds different meanings for respondents, particularly students who possess distinct belief systems and value orientations. Consequently, averaging scores for these items may yield misleading interpretations.

The third strategy hinges on a periodic evaluation of certain courses. Some university administrators contend that online evaluation technologies enable them to conduct evaluations for each course individually. Nevertheless, it is feasible to exempt certain groups of instructors from evaluations in particular semesters. For instance, distinguished or exemplary instructors, as well as full professors with outstanding educational records and research accomplishments. For these groups, it may be preferable that the SET items be proposed by the instructors themselves, allowing for targeted feedback aimed at enhancing their teaching practices.

References

Adams, G. L., & Engelmann, S. (1996). *Research on Direct Instruction: 25 Years Beyond DISTAR*. Seattle: Educational Achievement Systems.

Amini, A., Pishghadam, R. & Saboori, F. (2019). On the Role of Language Learners' Psychological Reactance, Teacher Stroke, and Teacher Success in the Iranian Context. *Journal of Research in Applied Linguistics*, *10*(2), 25-43. doi: 10.22055/rals.2019.14716

Borich, G. D. (2017). *Effective Teaching Methods: Research-Based Practice* (9th Ed.). Boston: Pearson Education.

Boswell, S. S. (2024). Academic entitlement and Ratemyprofessors.com evaluations bias student teaching evaluations: implications for faculty evaluation and policy-lenient professors' occupational health. *Heliyon*, *10*(8), 29-47. doi: https://doi.org/10.1016/j.heliyon.2024.e29473

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd Ed.). New York: The Guilford Press.

Buller, J. L. (2012). *Best practices in faculty evaluation: A Practical guide for academic leaders.* San Francisco. Jossey-Bass.

Burden, P. R., & Byrd, D. M. (2019). *Methods for Effective Teaching: Meeting the Needs of All Students*. (8th Ed.). New York: Pearson.

Burisch, M. (1998). Test Length and Validity Revisited. *European Journal of Personality*, *11*(4), 303-315. doi.org/10.1002/(SICI)1099-0984

Clift, J., Hall, C., & Turner, I. (1989). Establishing the Validity of a Set of Summative Teaching Performance Scales. *Assessment & Evaluation in Higher Education*, *14*(3), 193–206. doi.org/10.1080/0260293890140305

Cohen, R. J., Schneider, J. W., Tobin, R. M. (2022). *Psychological Testing and Assessment: An Introduction to Tests and Measurement* (10th Ed.). New York: McGraw Hill.

Cunningham, J. B., & MacGregor, J. N. (2006). The Echo Approach in Developing Items for Student Evaluation of Teaching Performance. *Teaching of Psychology*, *33*(2), 96–100. doi.org/10.1207/s15328023top3302_3

Dale, T., & Walsoe, H. (2020). Optimizing Grid Questions for Smartphones: A Comparison of Optimized and Non-Optimized Designs and Effects on Data Quality on Different Devices. In: Paul C. Beatty & et al. (Editors). *Advances in Questionnaire Design, Development, Evaluation, and Testing*. Hoboken, NJ: John Wiley & Sons Inc., 375-402. doi.org/10.1002/9781119263685.ch15

Dillman, D. A. (2020). Asking the Right Questions in the Right Way: Six Needed Changes in Questionnaire Evaluation and Testing Methods. In: Paul C. Beatty & et al. (Editors). *Advances in Questionnaire Design, Development, Evaluation, and Testing*. Hoboken, NJ: John Wiley & Sons Inc., 25-45. doi.org/10.1002/9781119263685.ch2

Dimitrov, D. M. (2011). *Statistical Methods for Validation of Assessment Scale Data in Counseling and Related Fields*. Alexandria: American Counseling Association.

Dodeen, H. (2013). Validity, Reliability, and Potential Bias of Short Forms of Students' Evaluation of Teaching: The Case of UAE University. *Educational Assessment*, *18*(4), 235–250. doi.org/10.1080/10627197.2013.846670

Dykema, J., Schaeffer, N. C., Garbarski, D., & Hout, M. (2020). The Role of Question Characteristics in Designing and Evaluating Survey Questions. In: Paul C. Beatty & et al. (Editors). *Advances in Questionnaire Design, Development, Evaluation, and Testing*. Hoboken, NJ: John Wiley & Sons Inc., 119-152. doi.org/10.1002/9781119263685.ch6

Finch, H. W., & French, B. F. (2019). *Educational and Psychological Measurement*. New York: Routledge.

Finch, H. W., Immekus, J. C., & French, B. F. (2016). *Applied Psychometrics Using SPSS and AMOS*. Charlotte, NC: Information Age Publishing, INC.

Fitzpatrick, A. R., & Yen, W. M. (2001) The Effects of Test Length and Sample Size on the Reliability and Equating of Tests Composed of Constructed-Response Items, *Applied Measurement in Education*, *14*(1), 31-57. doi.org/10.1207/S15324818AME1401_04

Garrido, L.E., Peñaló-Sánchez, D., Abreu-Mariot, C. et al. (2024). Cross-Cultural Validation of a Spanish-Language Version of the Composite Abuse Scale (Revised) – Short Form (CASR-SF). Journal of Family Violence, *39*(-), 1509–1523. doi.org/10.1007/s10896-023-00585-9

Gelber, Scott., M. (2020). *Grading the College: A History of Evaluating Teaching and Learning. Baltimore*. USA: Johns Hopkins University Press.

Gilbert, R. O., & Gilbert, D. R. (2025). Student evaluations of teaching do not reflect student learning: an observational study. *BMC medical education*, *25*(1), 313. https://doi.org/10.1186/s12909-025-06896-3

Gonzalez, O. (2021). Psychometric and Machine Learning Approaches to Reduce the Length of Scales. *Multivariate behavioral research*, *56*(6), 903–919. doi.org/10.1080/00273171.2020.1781585

Graf, P. (2024). Making Sense of Today's Use of Student Evaluations of Teaching (SET). *Human Arenas*, *7*(-), 446–450. doi.org/10.1007/s42087-023-00377-z

Gursoy, D., & Umbreit, W. T. (2005). Exploring Students' Evaluations of Teaching Effectiveness: What Factors are Important? Journal of Hospitality & Tourism Research, *29*(1), 91-109. doi.org/10.1177/1096348004268197

Hanh N. T. T. (2023). Organizational Factors Affecting Lecturer Performance in HEIs in the Context of Blended Learning: An Empirical Study in Vietnam. *johepal*. *4*(4), 32-49. doi:10.61186/johepal.4.4.32

Heffernan, T. (2021). Sexism, racism, prejudice, and bias: a literature review and synthesis of research surrounding student evaluations of courses and teaching. *Assessment & Evaluation in Higher Education*, *47*(1), 144–154. https://doi.org/10.1080/02602938.2021.1888075

Hodson, G. (2025). It is time to abandon student evaluations of teaching. *Nat Rev Psychol*, *4*, 433–434. https://doi.org/10.1038/s44159-025-00444-y

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. doi.org/10.1080/10705519909540118

Huang, TH., & Liu, YC. (2024). Evaluating Online Teaching Among Teachers, with an Emphasis on Scale Development and Validation Within Gender and Experience-Based Groups. *Asia-Pacific Educational Researcher*. doi.org/10.1007/s40299-024-00879-6

Kahn, P., & Anderson, L. (2019). *Developing your teaching: towards excellence* (2nd Ed.). New York: Routledge.

Kauchak, D. & Eggen, P. (2017). *Introduction to Teaching: Becoming A Professional* (6th Ed.). Boston: Pearson Education.

Khanjani, A., Vahdany, F. & Jafarigohar, M. (2017). EFL Teacher Education in Iran: Does It Promote Trainees' Pedagogical Content Knowledge? *Journal of Research in Applied Linguistics*, *8*(2), 159-186. doi: 10.22055/rals.2017.13096

Kim, H., Ku, B., Kim, J. Y., Park, Y. J., & Park, Y. B. (2016). Confirmatory and Exploratory Factor Analysis for Validating the Phlegm Pattern Questionnaire for Healthy Subjects. *Evidence-Based Complementary and Alternative Medicine*, 1-9, doi.org/10.1155/2016/2696019

King, H. (2022). *Developing exercises for teaching in higher education: Practical ideas for professional learning and development*. London: Routledge.

Kogar, H. (2020). Development of a Short Form: Methods, Examinations, and Recommendations. *Journal of Measurement and Evaluation in Education and Psychology*, *11*(3), 301-310. doi.org/10.21031/epod.739548

Kutscher, T. & Eid, M. (2020). The Effect of Rating Scale Length on the Occurrence of Inappropriate Category Use for the Assessment of Job Satisfaction: An Experimental Online Study. *Journal of Well-Being Assessment*, *4*(-), 1-35. doi.org/10.1007/s41543-020-00024-2

Lakeman, R., Coutts, R., Hutchinson, M., Lee, M., Massey, D., Nasrawi, D., & Fielden, J. (2021). Appearance, insults, allegations, blame and threats: an analysis of anonymous non-constructive student evaluation of teaching in Australia. *Assessment & Evaluation in Higher Education*, *47*(8), 1245–1258. doi.org/10.1080/02602938.2021.2012643

Landrum, R. E., & Braitman, K. A. (2008). The effect of decreasing response options on students' evaluation of instruction. *College Teaching*, *56*(4): 215-217.

Liebe, U., Glenk, K., Oehlmann, M., & Meyerhoff, J. (2015). Does the use of mobile devices (tablets and smartphones) affect survey quality and choice behaviour in web surveys? *Journal of Choice Modelling*, *14*(-), 17–31. doi.org/10.1016/j.jocm.2015.02.002

Linask, M., & Monks, J. (2018). Measuring faculty teaching effectiveness using conditional fixed effects. *The Journal of Economic Education*, *49*(4), 324–339. doi.org/10.1080/00220485.2018.1500957

Marks, R. B. (2000). Determinants of Student Evaluations of Global Measures of Instructor and Course Value. *Journal of Marketing Education*, *22*(2), 108-119. doi.org/10.1177/0273475300222005

Marsh, H. W., & Balla, J. (1994). Goodness of fit in confirmatory factor analysis: The effects of sample size and model parsimony. *Quality & Quantity*, *28*(2), 185–217. doi.org/10.1007/BF01102761

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist, 52*(11), 1187–1197. doi.org/10.1037/0003-066X.52.11.1187

McCoach, D. B., Gable, R. K., & Madura, J. P. (2013). *Instrument development in the affective domain: School and corporate applications* (3rd ed.). Springer New York, 1-307. doi.org/10.1007/978-1-4614-7135-6

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749. doi.org/10.1037/0003-066X.50.9.741

Mulaik, S.A. (2009). *Linear Causal Modeling with Structural Equations* (1st Ed.). Chapman and Hall/CRC, 1-468. doi.org/10.1201/9781439800393

Nederhand, M., Auer, J., Giesbers, B., Scheepers, A., & van der Gaag, E. (2022). Improving student participation in SET: effects of increased transparency on the use of student feedback in practice. *Assessment & Evaluation in Higher Education*, *48*(1), 107–120. doi.org/10.1080/02602938.2022.2052800

Niemand, T., & Mai, R. (2018). Flexible cutoff values for fit indices in the evaluation of structural equation models. *Journal of the Academy of Marketing Science*, *46*(-), 1148–1172. doi.org/10.1007/s11747-018-0602-9

Niemi, R. G., Carmines, E. G. & McIver, J. P. (1986). The impact of scale length on reliability and validity - A clarification of some misconceptions. *Quality and Quantity*, *1*(20), 371-376.

Park, B., & Cho, J. (2022). How does grade inflation affect student evaluation of teaching? *Assessment & Evaluation in Higher Education*, *48*(5), 723–735. doi.org/10.1080/02602938.2022.2126429

Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003). *Making sense of factor analysis*. SAGE Publications, Inc. doi.org/10.4135/9781412984898

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta psychologica*, *104*(1), 1–15. doi.org/10.1016/s0001-6918(99)00050-5

Rolph, K. E., King, A., Lardé, H., Kim, S. T., Ragland, N., Hervé Claude, L. P., French, H. M., & Gilbert, R. O. (2023). Critique of Approaches for Evaluating Teaching and Proposal for a New Institutional Policy. *Journal of veterinary medical education*, *50*(5), 499–507. https://doi.org/10.3138/jvme-2022-0072

Rolstad, S., Adler, J., & Ryden, A. (2011). Response Burden and Questionnaire Length: Is Shorter Better? A Review and Meta-analysis. *Value in Health*, *14*(8), 1101-1108. doi.org/10.1016/j.jval.2011.06.003

Sahragard, R. & Farangi, M. R. (2017). An Investigation of Curriculum Genre Using Rose (2014) Pedagogic Exchange Model. *Journal of Research in Applied Linguistics*, *8*(Proceedings of the Fourth International Conference on Language, Discourse and Pragmatics), 262-271. doi: 10.22055/rals.2017.12930

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting Structural Equation Modeling and Confirmatory Factor Analysis Results: A Review. *The Journal of Educational Research*, *99*(6), 323-337. doi.org/10.3200/JOER.99.6.323-338

Shrout, P. E., & Yager, T. J. (1989). Reliability and validity of screening scales: Effect of reducing scale length. *Journal of Clinical Epidemiology*, *42*(1), 69–78. doi.org/10.1016/0895-4356(89)90027-9

Slavin, R. E. (2018). *Educational Psychology: Theory and Practice* (12th Ed.). New York: Pearson.

Spooren, P., & Christiaens, W. (2017). I liked your course because I believe in (the power of) student evaluations of teaching (SET). Students' perceptions of a teaching evaluation process and their relationships with SET scores. *Studies in Educational Evaluation*, *54*(-), 43-49. doi.org/10.1016/j.stueduc.2016.12.003

Taherdoost, H. (2019). What Is the Best Response Scale for Survey and Questionnaire Design; Review of Different Lengths of Rating Scale / Attitude Scale / Likert Scale. *International Journal of Academic Research in Management*, *8*(1), 1-12.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC: American Psychological Association.

Uttl, B. (2024). Student Evaluation of Teaching (SET): Why the Emperor Has No Clothes and What We Should Do About It. *Human Arenas*, *7*(-), 403–437. doi.org/10.1007/s42087-023-00361-7

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

Watkins, M. W. (2021). *A step-by-step guide to exploratory factor analysis with SPSS*. New York: Routledge.

Willis, G. B. (2020). Questionnaire Design, Development, Evaluation, and Testing: Where Are We, and Where Are We Headed? In: Paul C. Beatty & et al. (Editors). *Advances in Questionnaire Design, Development, Evaluation, and Testing*. Hoboken, NJ: John Wiley & Sons Inc.

Xia, Y., Yang, Y. (2019). RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. *Behav Res*, *51*(-), 409–428. doi.org/10.3758/s13428-018-1055-2