Manipulating Grades and Evaluations: Unpacking the Power Dynamics Between Students and Teachers in Higher Education

Riaz Ahmad[1], Idrees Waris[2] and Muhammad Ayyoub[3]*

[1]Assistant Professor, Department of Management Sciences, University of Turbat, Turbat.Pakistan. Email: riaz.ahmed@uot.edu.pk (ORCID: 0000-0003-4861-8523).

[2]Assistant Professor, Department of Management Sciences, University of Turbat, Turbat.Pakistan. Email: idrees.waris@uot.edu.pk (ORCID: 0000-0002-7666-145X).

[3]*Corresponding Author: Assistant Professor, Department of Economics, University of Sahiwal, Sahiwal.Pakistan. Email: m.ayyoub@uosahiwal.edu.pk (ORCID: 0000-0001-5559-6188).

Journal of Educational Psychology and Pedagogical Sciences (JEPPS)

ISSN:2791-0393 (Print) eISSN: 2791-0407

Vol.5, No. 1, (Jan-June, 2025): 15-38

Submitted 20 March 2025

Accepted 5 June 2025

Published 30 June 2025

DOI: https://doi.org/10.52587/jepps.v5i1.113

https://jepps.su.edu.pk/article/46

OPEN ACCESS

ABSTRACT

The methodological challenges in evaluation of teaching and learning in higher education institutions (HEIs) worldwide emphasize the pressing need for reliable metrics that measure teaching quality and effectiveness. Despite concerns, HEIs still rely on Student Evaluation of Teaching (SET) scores on key faculty decisions (e.g. promotion). This study investigates the reciprocal dynamics between SET and TAL in HEIs in Pakistan. The research investigates whether SET and TAL scores can influence one another and whether these interactions contribute to grade inflation. Using a Manipulating Power Matrix that categorizes the dynamics of manipulative power between teachers and students, this study examines the correlation between SET scores and TAL results across various assessment components including sessional assessment (i.e. formative assessment) and midterm and final term examinations (i.e. summative assessments). Ordered logistic regression on 6,230 SET responses (across 297 courses) shows a strong positive link between SET scores and sessional assessment results with little to no correlation in midterm or final exams. The sessional component—most prone to manipulation—seemingly appears to drive both higher SET ratings and grade inflation. The findings of the study question the use of SET scores for evaluation of teaching quality and highlight the impact of institutional policies on outcomes in HEIs Pakistan. We recommend a more balanced approach that can integrate the SET scores with peer reviews and other teaching portfolios. This triangulated approach not only mitigates the risk of grade inflation but it also ensures a fairer and accurate evaluation process.

Keywords: Higher Education; Students' Evaluation of Teaching (SET); Teachers' Assessment of Learning (TAL); Grade Inflation; Summative Assessment; Formative Assessment.

1. Introduction

Student Evaluations of Teaching (SET) surveys have gotten much attention in academia though the reliability and validity of the surveys remained in question. For example, Uttl (2024) criticizes the SET scores fail to predict teaching performance and relying on this only for career-

related decisions in particular is flawed. similarly, Graf (2024) acknowledges the limitations of SET in measuring teaching quality, though it is valid indicators for student satisfaction.

In the realm of higher education, the longstanding tradition of employing Students' Evaluation of Teaching (SET) scores to evaluate the effectiveness of faculty members has exerted a profound influence on various facets of academic life. Remarkably, an overwhelming 86% of higher education institutions (HEIs) depend on SETs as a crucial criterion in shaping personnel decisions (Seldin, 1993). Examining 1139 documents spanning the period from 1964 to 2018, co-occurrence analysis reveals a discourse on the implementation of Students' teaching evaluation (i.e. SET), indicating its emergence in the US during the 1970s, spreading to German-speaking countries in the mid-1990s, and later reaching China and Latin America in the early 2000s (Pineda & Steinhardt, 2023).

Despite acknowledged limitations, encompassing validity concerns and biases linked to halo effects, gender, race, and prejudice (Binderkrantz, *et al.,* 2022; Cannon & Cipriani, 2022; Heffernan, 2022; Michela, 2023; Sigurdardottir, *et al.,* 2023), SET surveys persist as the primary tool for assessing teaching quality and motivating educators. While widely regarded as valid instruments for evaluating teaching practices, recent meta-analyses have cast doubt on this assumption (Uttl et al., 2017). The discourse surrounding SET scores has become a focal point of extensive research, yielding conflicting views on their efficacy in gauging teaching quality and teacher expertise (Cannon & Cipriani, 2022; Spooren, *et al.,* 2013; Uttl *et al.,* 2017; Wright & Jenkins-Guarnieri, 2012).

Furthermore, SET scores have been demonstrated to influence the grades students receive from instructors, potentially leading to grade inflation, as instructors may be inclined to award higher grades to students providing favourable evaluations (Jin, 2019; Johnson, 2006; Krautmann & Sander, 1999; Weinberg, *et al.,* 2009). The reciprocal relationship in the teaching of higher education setting poses a challenge in terms of compromising instructional goals and teaching effectiveness (He et al., 2022). In addition, inconsistencies have also been noted in association of SET to student learning; with a strong and positive association found between SET and practical examination scores but no significant correlation with multiple-choice test scores (Stehle, et al., 2012).

In addition to the above concerns of biasness in SET surveys, the methodological challenges in association to data analysis of SET surveys is also another area of concerns that led researchers to formulate and employ more sophisticated analytical tools and approaches. In this regard, Huang and Cai (2023) introduced a model of cross-classified Item Response Theory (IRT) that can not only handle the multilevel and hierarchical nature of SET data but also offers a more precise method for analysing data of student feedback. The IRT model also overcomes the traditional limitations of the parametric approaches which have often faced criticism for their inclination toward oversimplify complex evaluative data. These scholars caution against the extensive use of parametric methods for the purpose of analysing Likert-scale data from SET surveys by highlighting the potential risk of deriving invalid conclusions that may unduly affect faculty performance evaluations.

The impact of SET scores on teacher is a serious issue that is well highlighted in the academic literature. As Hodson (2025) and Gatwiri, et al. (2024) argues, SET surveys often discriminate women and racialized groups that lack both construct and predictive validity. This concern is further echoed by Hutchinson et al. (2024), who explored how anonymous and often non-constructive feedback can affect the well-being of younger and untenured faculty members negatively. They observed that the SET's findings, if shared, can create a hostile environment for instructors by affecting their professional confidence and mental health. This body of literature

raises very important concerns, ethical or methodological, about the role of SETs in perpetuating inequities and usage within HEIs.

Even in context of Pakistan, research on SET surveys and their impact on quality of education remains limited but crucial, given that a number of unique challenges faced by HEIs in the country. In this regard, Wahid, *et al.* (2023) provide a qualitative analysis of SET practices in Pakistan that highlights several key issues including, but not limited to, the connection between grade inflation and SET scores, low response rates, the influence of teacher characteristics on student perceptions, etc. While SETs are regarded as essential instruments for quality assurance; their efficacy is compromised by biases associated with the physical appearance of instructors as well as by the propensity of visiting faculty members to inflate students'' grades in order to secure more favourable scores in these evaluations. Moreover, Ikram & Kenayathulla (2023) explored how quality of education influences student satisfaction in private universities in Pakistan. They found a positive association between students' perceptions and improvements in instructional materials, support and classroom facilities. However, they also noted that the overemphasis on student satisfaction can detract the policymakers or academic community from more objective measures of educational quality.

The gaps in the literature become evident when examining the interplay between SET and Teachers' Assessment of Learning (TAL), predominantly in contexts such as Pakistan where these assessments are often viewed in isolation. Existing studies primarily focus on the impact of SET on teaching quality and its methodological and ethical challenges but seldom explore the bidirectional relationship between student evaluations and teacher assessments. So, this study addresses this gap by investigating the correlation between SET and TAL scores, with a focus on the potential for manipulation by both students and teachers. The introduction of a Manipulating Power Matrix provides a novel framework for understanding power dynamics in the evaluation process that offers insights into how sessional assessments (a component in overall assessment), in particular, may influence student and teacher behaviour.

Despite having extensive research available on the SET scores, the critical issues still persist without any concrete resolution that motivate this analysis. In this paper, it was explored potential manipulation within the context of assessment and evaluation caused from the reciprocal relationship between SET scores and student grades. In this study we examined the impact of the assessment policies by the Higher Education Commission (HEC), Pakistan on Students' Evaluation of Teaching (SET) and Teachers' Assessment of Learning (TAL) focusing on a public sector university in Pakistan, namely, the University of Turbat. The research further investigates into the components of TAL (i.e. the sessional, midterm exams, and final term exams) in order to comprehend their association with the SET scores. It is noteworthy to clarify the term of assessment and evaluation here; in this context, 'assessment' refers to the assessment of student learning through their examinations' scores while 'evaluation' denotes Students' Evaluation of Teaching through the surveys scores.

This study contributes significantly to existing literature in various ways. First, the central theoretical contribution of this study lies in the development of the 'Power Matrix of Manipulation in SET and TAL: Teachers vs Students'. This matrix provides a novel conceptual framework to understand and categorize the dynamics of manipulation within the processes of Students' Evaluation of Teaching (SET) and Teachers' Assessment of Learning (TAL). The framework is grounded in the idea that both students and teachers can exert varying degrees of influence over one another, particularly in the context of different types of academic assessments. This manipulation can lead to skewed SET scores and grade inflation, thereby compromising the validity and reliability of these metrics as indicators of teaching quality and student learning.

Second, we tried to identify a specific area of assessment in which teachers and students may influence each other for securing higher SET scores and grade points, respectively. Unlike others, we use students' assessment with the components (i.e. sessional, midterm exam, and final term exam) in percentage for identifying the specific component through which this potential biases in SET scores or grade inflation might occur due to the reciprocity between students and teachers.

Third, compliance to HEC's quality standards, the university's policies on grading, examination timing, SET surveys and the time of results dissemination provide this study a robust framework to investigate the relationship between SET and TAL. The variation in time of conducting and disseminating TAL and SET scores may play huge role in variation of SET scores (see Table 4 for details). For instance, if SET surveys were conducted when students received their final grades, students inclined to rate their teachers based on their relative grades in exams. Also, instructors often award higher grades to students with better SET scores when SET surveys conducted before the announcement of exam results (Berezvai, et al. 2021).

Last, we analysed each SET item (i.e. the questions) separately and those were categorized into groups based on their thematic association: Teacher's Punctuality, Teacher's Seriousness, Teacher's Personal Characteristics, and Teacher's Subject Knowledge.

In this study, these two fundamental research questions were addressed:

1. Is there a relationship between Students' Evaluation of Teaching (SET) and Teachers' Assessment of Learning (TAL)? If so, what kind of relationship is it?

2. Do students and teachers utilize the TAL or SET tools to manipulate each other for achieving higher scores on their respective report cards? If so, which aspects of assessment or evaluation contribute to such manipulation more?

This study tries to answer these questions by providing a unique perspective on the reciprocal influence of student evaluations and teacher assessments in the higher education settings. It also sheds light on potential biases and manipulations that may happen during the evaluation or assessment processes.

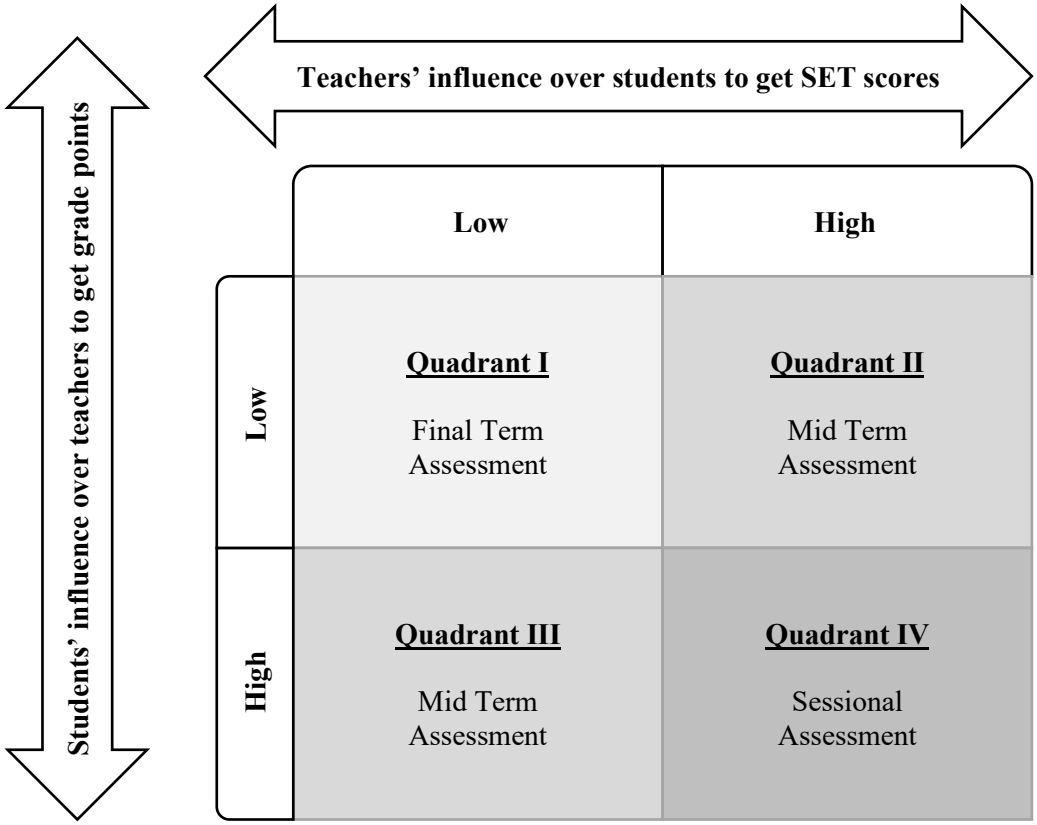
2. Conceptual Framework and Hypotheses Development

The research was conducted within the operational framework of the Higher Education Commission of Pakistan, functioning as the regulatory and funding body for Higher Education Institutes (HEIs). The focal point of this investigation was a public sector institution operating under the umbrella of HEC, Pakistan.

The conceptual framework of this study is built on connection between SET and TAL in context of HEIs in Pakistan in which we frame a model that explain the power dynamics—between students and teachers. We construct a Matrix of Power Manipulation that categorizes the power dynamics into four quadrants (see Figure 1).

Figure 1

Power Matrix of Manipulation in SET and TAL: Teachers vs Students



In this study, we use the term '**quadrant' which** refers to each of the **four distinct sections** within the above *Power Matrix* that is a conceptual framework developed for mapping and analyzing the varying degrees of influence and control exercised by students and teachers over academic evaluations. Here each **quadrant** represents a unique combination of:

a. **Student power** (through SET)
b. **Teacher power** (through TAL)
c. **Timing** of assessments relative to evaluations
d. **Institutional control** mechanisms

Hypothesis 1

H1: There is a positive and significant correlation between Students' Evaluation of Teaching (SET) scores and their overall academic performance.

This hypothesis asserts a positive correlation between students' perceptions of teaching effectiveness (SET scores) and their actual performance in assessments. Utilizing overall percentage marks and their individual components, the study aims to pinpoint specific assessment elements influencing the teaching and learning process, thereby identifying potential SET score biases. The choice of total marks as a measure offers advantages over GPA, including a robust correlation with GPA, increased variability, and the ability to dissect assessments into their components. The hypothesis anticipates that higher SET scores align with superior student performance.

The second research objective guides the study on the reciprocal manipulation of SET and TAL by students and teachers respectively that is operationalized through the analysis of how each of the assessment component relates to a given SET score. We assume that a stronger and statistically significant correlation between SET scores and a specific component (e.g., sessional assessment) would indicate an area for higher potential in manipulation. It would be more evident in component which is more flexible or instructor-controlled. So, the Hypothesis 2 given below is designed not only to test the presence and strength of such types of relationships but also to identify the area or components of assessment in which reciprocal power dynamics are most likely to occur.

Hypothesis 2

H2: The relationship between SET scores and students' academic performance across various components (i.e. sessional, midterm and final examinations) varies in both strength and direction. More specifically:

The correlation between SET scores and sessional examination marks (in %) is positive and statistically significant.

The association between SET scores and midterm examination marks (in %) is ambiguous or indeterminate.

The correlation between SET scores and final examination marks (in %) is weak or statistically insignificant.

This hypothesis recognizes the intricacies in power dynamics during midterm assessments (Quadrant II and Quadrant III). Depending on when midterm results are disclosed concerning the SET survey, the relationship between SET scores and midterm exam marks can vary from positive to negative or remain neutral. The timing of midterm exams and SET surveys creates two indeterminate situations, acknowledging limited power for manipulation due to institutional policies. The hypothesis aims to explore the unpredictable nature of the relationship in these specific conditions. Additionally, it focuses on Quadrant IV, where both teachers and students possess significant manipulative power. The reciprocal manipulation, in this quadrant, seemingly causes a positive relationship between SET scores and sessional marks. The study posits that the dynamics of continuous assessments in the sessional component such as assignments, quizzes, presentations, etc. create such an environment in which both parties (i.e. students and teachers) can leverage their power for mutual benefit.

In addition, this relationship between SET and TAL in Quadrant I in the matrix is weak that emphasizes that the final term assessment (final exams) is assumed to be a better proxy among others to measure the quality of teaching and learning. It is because of the reason that, due to external (by HEC or accreditation councils) and internal (by QEC) monitoring and evaluation, teachers most probably to prefer to design their final exam papers rigorously that would be aligned with Intended Learning Outcomes (ILOs) rather than focusing on manipulation for the sake of better SET scores. Therefore, we hypothesize in this study that there is no or weak correlation between SET scores and final term exam marks.

3. Research Setting

The research was conducted within the operational framework of the Higher Education Commission of Pakistan which is functioning as the regulatory (or monitoring) and funding body for Higher Education Institutes (HEIs) in Pakistan. The research setting of this study is the University of Turbat (UoT), a public sector university in Pakistan. The university enrols about 4000 students, with more than 45% of females. This study was conducted in 2021 that targeted students enrolled in thirteen undergraduate degree programs.

Like other universities in Pakistan, the Quality Enhancement Cell (QEC) oversees the administration, analysis and dissemination of SET scores to course instructors. Adhering to institutional policies, the survey was conducted two weeks before final term examinations, with SET scorecards sent to instructors post the overall examination results announcement (refer to Table 1). In terms of TAL, measured in marks in percentage, the university's semester regulations divide it into three components: sessional (30% marks), midterm exam (30% marks), and final term exam (40% marks). Unlike traditional methods that rely on Grade Point Average (GPA), this study emphasizes total marks and individual TAL components. This distinction is essential for three reasons: it is a GPA substitute with a 97% correlation, it allows for greater variability in assessment marks compared to GPA, and it enables the investigation to identify specific assessment components susceptible to manipulation by teachers and students.

The research confronts the challenges of bias and validity in SET scores, presuming these issues stem from the dual nature of assessment and evaluation processes. The institutional policies governing assessment and evaluation at the university offer a structured research design. The investigation delves into the power dynamics between teachers and students, exploring how each party may influence the other to achieve higher SET scores and grade points, respectively. The study dissects the temporal dynamics of assessment components, focusing on sessional marks, midterm exam marks, and final term exam marks. Sessional, encompassing various activities throughout the semester, is at the discretion of course instructors, contributing 30% to the overall assessment. Midterm exam, a 1-2 hours written paper, occurs in the 8th week, while the final term exam, a comprehensive 2-3 hours written paper, concludes the semester in the sixteenth and last week of the semester (refer to Table 1 and Figure 2).

Table 1

Schedules of Assessment, Evaluation and Dissemination of Results

Activity	Occurrence During the Semester
Assessment: Sessional	1st – 16th weak
Assessment: Mid Term Exam	8th week
Assessment: Final Term Exam	16th week
Evaluation: SET Survey	14th week
Assessment: Results Announcement	17th - 18th week
Evaluation: SET Scores Dissemination	19th - 20th week

While GPA is a consolidated measure in traditional assessments, this study values the granularity of individual components. Notably, Sessional marks offer a platform for reciprocal manipulation between teachers and students, leading to a hypothesized positive correlation between SET scores and Sessional marks. The final term exam, with its 40% weightage, emerges as a critical measure of teaching and learning quality due to its comprehensive nature, alignment with Intended Learning Outcomes (ILOs), and is subject to scrutiny from internal (e.g. QEC) and external bodies (e.g. HEC). The research setting at the university provides a nuanced lens to examine the intricate relationships between SET and TAL, offering insights into the dynamics of assessment practices and their implications on teaching and learning outcomes.

Figure 2

The Conceptual Framework – Timing

		Week																	
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17-18	19-20
Teacher Assessment of Learning (TAL)																			
Sessional (30%)	Conduct	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●		
	Results Dissemination	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●		
Mid Term (30%)	Conduct								●										
	Results Dissemination									●	●	●	●	●	●	●	●	●	
Final Term (40%)	Conduct																●		
	Results Dissemination																	●	
Student Evaluation of Teaching (SET)																			
Survey	Conduct														●				
	Score Dissemination																		●

4. Methods/Methodology

Research Design

In this study, we use a research design in quantitative and explanatory nature to examine the association between SET and TAL in HEIs in Pakistan. This research integrates secondary data of students' assessment scores with the student survey responses. We also use a conceptual framework—'Power Matrix of Manipulation in SET and TAL'—to guides the process of operationalization of variables and interpretation findings.

Survey Instrument and Data Collection

We used thirteen items (i.e. questions) from the HEC-designed SET survey questionnaire (see Batool & Qureshi, 2007, for the questionnaire). The survey was conducted by QEC two weeks before the final term exams in Fall 2021. Trained enumerators administered the survey in a controlled environment similar to a regular classroom setting. The survey achieved a response rate of approximately 95 percent. At the time of data collection, the total undergraduate student population at the University was 3,414. From this population, a sample of 6,230 SET responses was collected from students enrolled in 297 courses across thirteen undergraduate degree programs.

Semester Results Data

Student assessment information, including students' grade points average (GPA), total marks, sessional marks, midterm marks, and final term examination marks, along with class size, semester, degree program, and other control variables, were obtained from the university's QEC and Controller of Examinations offices.

Ethical Considerations

This study adhered to rigorous ethical protocols to ensure the confidentiality and privacy of all participants. The primary data were derived from the QEC's surveys (i.e. teacher and course. .

evaluations surveys), in which neither the names of students nor their identification numbers were recorded. However, course identifiers were included for analytical purposes only. The second data source comprised aggregated examination results that reflect the class averages for specific courses rather than individual student scores. In order to maintain anonymity, unidentified student evaluation responses were matched with the corresponding course's average assessment scores (i.e. results). All data were systematically coded and anonymized to ensure that neither individual instructors nor students could be identified. A written approval for the utilization of data was obtained from the Vice Chancellor of the University. This consideration aligns with our ethical commitment to preserve the privacy of participants.

5. Descriptive Statistics

Table 2 outlines the descriptive statistics for Teachers' Assessment of Learning (TAL) variables, encompassing grade points average (GPA), total marks (percentage), sessional marks (percentage), midterm exam marks (percentage), and final term exam marks (percentage). The data, derived from 5,787 students across various faculties and semesters, portrays a comprehensive overview. The mean GPA is 3.36, with a standard deviation of 0.41, suggesting a relatively narrow distribution around the mean within the [1.00 – 4.00] scale. Total Marks (percentage) exhibit a mean of 77.33 and a standard deviation of 6.25, signifying variability in overall performance on a [0 – 100] scale.

Sessional marks (percentage) have a mean of 24.41 and a standard deviation of 2.41, while midterm exam marks (percentage) and final term exam marks (percentage) have means of 22.39 and 30.56, respectively, with corresponding standard deviations of 2.70 and 3.53. These statistics offer insights into the distribution of marks within specific assessment components. An average class size of 47 students (with a S.D. of 33) indicates a substantial variation in class sizes across courses——ranging from smaller classes to larger ones. Moreover, the data is broken down by faculty, with Economics, Commerce, and Business Administration constitute 19.11% of the sample, Science and Engineering make up 29.13%, Social Science, Arts, and Humanities encompass 41.82%, and Legal Education accounts for 9.94%. Semester-wise distribution is provided with the Second Semester comprising 44.98%, Fourth Semester at 19.75%, Sixth Semester at 30.67%, and Ninth Semester at 4.60%.

Table 2

Descriptive Statistics of Teachers' Assessment of Learning (TAL) Variables

Teachers' Assessment of Learning (TAL)		**Observation**	**Mean**	**Std. Dev.**
Grade Points Average (GPA)	[1.00 – 4.00]	5,758	3.36	0.41
Total Marks (%)	[0 – 100]	5,787	77.33	6.25
Sessional Marks (%)	[0 – 30]	5,787	24.41	2.41
Midterm Exam. Marks (%)	[0 – 30]	5,787	22.39	2.70
Final Term Exam. Marks (%)	[0 – 40]	5,787	30.56	3.53
Other Variable				
Class Size (Number of students)		5,787	47	33

Faculty	**Observation**	**Percent**
Economics, Commerce and Business Administration	1,106	19.11
Science and Engineering	1,686	29.13
Social Science, Arts, and Humanities	2,420	41.82
Legal Education	575	9.94
Semester	**Observation**	**Percent**
Second	2,603	44.98
Fourth	1,143	19.75
Sixth	1,775	30.67
Ninth	266	4.60

Table 3 presents the descriptive statistics for Students' Evaluation of Teaching (SET) across various outcome variables. The data is derived from a comprehensive assessment of teaching quality, with a focus on punctuality and attendance, preparation and engagement, student interaction and support and classroom management and fairness. The mean scores—based on a Likert-type scale of 1 to 3 (where, 1= disagree, 2=neutral and 3= agree)—reflect a general positive outlook from students across all outcome variables of interest. For example, the percentages in the column of "Students' Response" provide an understanding of the distribution of perceptions of students with a vast majority of them express positive agreement with the quality of teaching.

In category of *punctuality and attendance*, most students agreed that their instructors arrived and departed the classroom on time (mean score is 2.7) and maintained their regular class attendance (75–81% agreement). Students also reported a positive perception on their *instructors' preparation* and *subject knowledge*. In terms of *student interaction and support*, students generally agreed that their instructors addressed students concerns, encouraged class participation and were always available for consultation (mean score = 2.5–2.7 with agreement of 66–81%). In terms of *classroom management* and *fairness* in exams, students perceived that their instructors as fair in assessments and at the same time effective in creating a conducive environment for learning (mean = 2.5–2.6; agreement = 69–76%).

The descriptive statistics in Table 3 indicate a predominantly positive assessment of teaching quality, with students expressing favourable perceptions across various dimensions. These findings serve as a foundational understanding for further analysis and interpretation of the SET data.

Table 3

Descriptive Statistics of Students' Evaluation of Teaching (SET) for Outcome Variables

Outcome Variables: Students' Evaluation of Teaching (SET)	Obs.	Mean	Students' Response		
			Agree	Neutral	Disagree
Punctuality and Attendance					
Arrived on Time	5515	2.7	81%	9%	10%
Left on Time	5479	2.7	78%	10%	11%
Was Regular During the semester	5362	2.6	75%	12%	13%

Preparation and Engagement					
Prepared for the class	5548	2.7	83%	8%	9%
Completed the course	5482	2.6	77%	11%	13%
Demonstrated subject knowledge	5537	2.7	80%	11%	10%
Presented concepts clearly	5479	2.6	77%	11%	12%
Communicated effectively	5473	2.6	73%	13%	14%
Student Interaction and Support					
Dealt with students' problems	5454	2.5	66%	20%	14%
Respected and encouraged class participation	5548	2.7	81%	9%	10%
Was available for consultation	5557	2.7	78%	12%	10%
Classroom Management and Fairness					
Maintained a conducive learning environment	5518	2.6	76%	13%	11%
Was fair in examinations	5433	2.5	69%	16%	15%

Note. Mean scores are based on a scale of 1 to 3, where 1 represents disagree, 2 represents neutral, and 3 represents agree. Percentages in the "Students' Response" columns indicate the proportion of students falling into each category.

Correlation Analysis

Table 4 illustrates that there exist a statistically significant positive association between total marks (in %) and grade point average (GPA) with a correlation coefficient of 0.97. GPA shows a positive correlation with sessional (0.52), midterm (0.70) and final term exam marks (0.81). It is also observed that there also exists a correlation between total marks with its components: sessional (0.59), midterm (0.69) and final term (0.83). In contrast, a weaker correlation exists among the components with sessional-midterm (0.16), sessional-final term (0.25) and final-midterm (0.36). We perform two separate regression analyses: one for total marks and the other for its components.

Table 4

Correlation Among Components of Assessment

Variables	1	2	3	4	5
1. Grade Points Average	1.00				
2. Total Marks (%)	0.97	1.00			
3. Sessional Marks (%)	0.52	0.59	1.00		
4. Mid Term Marks (%)	0.70	0.69	0.16	1.00	
5. Final Term Marks (%)	0.81	0.83	0.25	0.36	1.00

Note. Authors' own calculation based on assessment data at the office of controller examinations of the university.

Estimation Model

We employed a series of ordered logistic regression analyses to investigate the association .

between the SET Scores on individual items and TAL scores, encompassing both overall and specific components (sessional, midterm, and final term exams marks). Ordered logistic regression is a regression model tailored for ordinal dependent variables (Fullerton, 2009; Harrell, 2001). In contrast to linear regression, which suits continuous dependent variables, and logistic regression, which suits dichotomous dependent variables, ordered logistic regression is well-suited for ordinal dependent variable s, as highlighted by (Hosmer, *et al.,* 2013). Following (Liu, *et al*., 2019), ordered logistic regression was employed in this study due to the ordinal nature of the dependent variable, which consists of three ordered categories representing Students' Evaluation of Teaching (*SET*): "Disagree" (P_1), "Neutral" (P_2), or "Agree" (P_3). The corresponding model is expressed by the following equation:

$$Logit\ (P_1) = log\frac{P_1}{1-P_2} = \beta' X + \varepsilon$$

$$Logit\ (P_1 + P_2) = log\frac{P_1 + P_2}{1-P_1-P_2} = \beta' X + \varepsilon$$

$$Logit\ (P_1 + P_2 + P_3\ ) = log\frac{P_1 + P_2 + P_3}{1-P_1-P_2-P_3} = \beta' X + \varepsilon$$

X represents a set of independent variables, ε denotes the error term, and β′ corresponds to the regression coefficients associated with the independent variables. The segment of the regression equation corresponding to β′X + ε can be expressed as follows:

$$\boldsymbol{SET_{ic}} = \boldsymbol{\alpha_0} + \boldsymbol{\alpha_1 TAL_c} + \boldsymbol{\alpha_2}\ \text{Size}_c + \sum \boldsymbol{\gamma_j}\ \text{T}_j + \sum \boldsymbol{\delta_k}\text{P}_k + \sum \boldsymbol{\varphi_l}\text{S}_l\ + \boldsymbol{\varepsilon_{icjkl}}$$

where SETic refers to the individual Students' Evaluation of Teaching (SET) scores for the i^{th} student taught by a teacher in course *c*, and TAL_c represents the course's class average of Teachers' Assessment of Learning (*TAL*) measured in total marks (as a percentage), the analysis incorporates fixed effects for Teacher designation (T_j), degree program (P_k), and semester (S_l). These fixed effects serve to account for variations at the instructor's designation level (professor, lecturer, visiting faculty), program level (BS Economics, BBA, BS Biotechnology, etc.), and semester level (2^{nd}, 4^{th}, 6^{th}, and 9^{th}). Additionally, Size_c is included to address the effects of class size on SET scores, drawing on previous studies (Andrade & Rocha, 2012; Berezvai *et al.*, 2021; Cho & Cho, 2017; Ewing, 2012). To further investigate the correlation between SET scores and TAL, the TAL is dissected into its assessment components in the following manner:

$$\boldsymbol{SET_{ic}} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}\text{SMc} + \boldsymbol{\beta_2}\text{MMc} + \boldsymbol{\beta_3}\text{FMc} + \boldsymbol{\beta_4}\text{Sizec} + \sum \boldsymbol{\gamma_j}\text{Tj} + \sum \boldsymbol{\delta_k}\text{Pk} + \sum \boldsymbol{\varphi_l}\text{Sl}\ + \boldsymbol{\varepsilon_{icjkl.}}$$

Here, SMc, MMc, and FMc denote the average scores for sessional, midterm, and final term assessment in course c. Table 4 indicates that the correlation among the independent variables of interest is quite weak, enabling the execution of the regression analysis. The ordinal logit model was utilized for estimation, and the obtained coefficients, along with marginal effects computed using STATA 13, are presented in Tables 5-8. Ordered logistic regression was employed due to the ordinal nature of SET responses (i.e., Disagree, Neutral, Agree). This model is more appropriate than linear regression in capturing the probability distribution of ranked outcomes and accounts for the ordered structure of the dependent variable. The application of this model is appropriate under the given assumptions: (a) the dependent variable is ordinal, (b) the observations are independent, (c) there is a linear relationship between continuous predictors and the log-odds of the outcome, (d) the model does not suffer from multicollinearity among predictors and (e) the sample size is sufficiently large to ensure statistical power and stable estimates. Each of these assumptions was carefully considered and met in the design and implementation of the analysis.

The outcomes of the ordered logit regression model, which explores the association between Teachers' Assessment of Learning (TAL) marks and Students' Evaluation of Teaching (SET) scores, are detailed in Tables 5-8. The dependent variables were evaluated using a Likert scale featuring categories such as disagree, neutral, and agree. The analysis comprises two panels: Panel A (total marks in percentage) and Panel B (components of total marks, including sessional marks, midterm marks, and final term marks in percentage), investigating the impact of diverse academic metrics on SET across various items or questions.

Instructor Punctuality and Attendance

In Panel A, Table 5, for the outcome variable "Arrive on Time," the coefficient for total marks in percentage is 0.0155 with a significance level of 0.05. This suggests that a one-unit increase in total marks is associated with a 0.0155-unit increase in the likelihood of students agreeing that the instructor arrives on time. The marginal effects reveal that higher total marks correspond to a decreased likelihood of disagreement (-0.0012) and neutral response (-0.0011), but an increased likelihood of agreement (0.0023).

Moving to Panel B, the analysis dissects the total marks into components. For "instructor arrived on time," the sessional marks component has the highest coefficient of 0.0709, indicating a stronger positive association with students agreeing that the instructor arrived on time compared to the other components. This finding aligns with the argument by Uttl (2024), who cautions against using SET scores for high-stakes decisions, as SETs do not necessarily reflect teacher effectiveness. However, Graf (2024) counters by asserting that SET scores validly measure student satisfaction with their instructors, as evidenced by the strong association between punctuality and sessional marks. The marginal effects further highlight the impact of sessional marks on reducing disagreement (-0.0055) and neutral responses (-0.0049), while increasing the likelihood of agreement (0.0104). This supports Park & Cho's (2022) conclusion that sessional assessments, as part of SET scores, can reflect students' immediate classroom experiences, which contributes to grade inflation. Similar patterns are observed for the "instructor left on time" and "instructor was regular during the semester" outcomes. In both cases, the sessional marks component consistently exhibits the highest coefficient in Panel B, suggesting its significant influence on students' perceptions.

The findings support H1, indicating a positive correlation between SET scores and students' overall academic performance. Regarding H2, the association between SET scores and different exam components varies. Sessional exams show a consistently strong positive correlation, midterm exams exhibit a somewhat ambiguous relationship, and final exams display a weaker or non-existent association relative to other components.

Table 5

Ordered Logit Regression Model Results: Punctuality and Attendance

				Marginal effects		
Outcome Variable: Instructor	**Independent Variable:** Class Marks (in %)	**Obs.**	**Coef.**	**Disagree**	**Neutral**	**Agree**
Panel A						
Arrived on Time	Total	5515	0.0155*	-0.0012*	-0.0011*	0.0023*
Left on Time		5479	0.0117	-0.0011	-0.0008	0.0019
Was Regular During the semester		5363	0.0188**	-0.0020**	-0.0014**	0.0034**
Panel B						
Arrived on Time	Sessional	5515	0.0709**	-0.0055**	-0.0049**	0.0104**
	Mid Term		-0.0534**	0.0041**	0.0037**	-0.0078**
	Final Term		0.0503**	-0.0039**	-0.0035**	0.0074**
Left on Time	Sessional	5479	0.0680**	-0.0065**	-0.0047**	0.0112**
	Mid Term		-0.0428**	0.0041**	0.0030**	-0.0071**
	Final Term		0.0352**	-0.0034**	-0.0024**	0.0058**
Was Regular During the semester	Sessional	5363	0.0665**	-0.0069**	-0.0051**	0.0121**
	Mid Term		-0.0113	0.0012	0.0009	-0.0020
	Final Term		0.0227	-0.0023	-0.0018	0.0041

Note. Standard errors are in parentheses. ** p<.01, * p<.05. The model is estimated using ordered logit model estimation techniques, with additional control variables such as class size as well as teacher designation, degree program, and semester fixed effects. Obs. and Coef. stand for Observations and Coefficients respectively. Authors' own calculation based on QEC surveys and assessment data at the university.

Instructor Preparation for Class and Engagement with Students

In Panel A, Table 6, the total marks variable exhibits a positive and statistically significant coefficient of 0.0366. This implies that a one-unit increase in total marks is associated with a 0.0366 increase in the odds of students agreeing with the statement that the instructor was prepared for the class. The marginal effects further reveal that higher total marks are associated with a 0.0027 decrease in the odds of disagreement, a 0.0021 decrease in the odds of a neutral response, and a 0.0048 increase in the odds of agreement.

Moving to the components in Panel B, the analysis is disintegrated into sessional, midterm, and final term exam marks. For the variable "instructor prepared for the class," the sessional marks have the highest coefficient (0.0921), indicating that performance in sessional exams has the most substantial impact on students' perception of instructor preparedness. These findings reinforce

Sullivan et al. (2024), who emphasize that SET surveys can lead to improvements in teaching, particularly when students perceive value in their participation. The marginal effects values for sessional, midterm, and final term further elucidate that sessional marks component has the greatest influence on the odds of agreement. Similar patterns are observed in the analysis of other items. For instance, in the case of "instructor presented concepts clearly," the sessional marks component again exhibits the highest coefficient (0.0798), emphasizing the critical role of sessional performance in influencing students' perceptions. The marginal effects values corroborate this, indicating that sessional marks component has the most pronounced impact on the odds of agreement.

Table 6

Ordered Logit Regression Model Results: Preparation and Engagement

				Marginal effects		
Outcome Variable: Instructor	**Independent Variable:** Class Marks (in %)	**Obs.**	**Coef.**	**Disagree**	**Neutral**	**Agree**
Panel A						
Prepared for the class	Total	5548	0.0366**	-0.0027**	-0.0021**	0.0048**
Completed the course		5482	0.0118	-0.0012	-0.0008	0.0020
Demonstrated subject knowledge		5537	0.0204**	-0.0016**	-0.0016**	0.0032**
Presented concepts clearly		5479	0.0405**	-0.0041**	-0.0029**	0.0069**
Communicated effectively		5473	0.0374**	-0.0042**	-0.0030**	0.0072**
Panel B						
Prepared for the class	Sessional	5548	0.0921**	-0.0069**	-0.0052**	0.0121**
	Mid Term		0.0349*	-0.0026*	-0.0020*	0.0046*
	Final Term		0.0135	-0.0010	-0.0008	0.0018
Completed the course	Sessional	5482	0.0554**	-0.0054**	-0.0041**	0.0094**
	Mid Term		-0.0128	0.0013	0.0009	-0.0022
	Final Term		0.0125	-0.0012	-0.0009	0.0021
Demonstrated subject knowledge	Sessional	5537	0.0516**	-0.0041**	-0.0039**	0.0080**
	Mid Term		0.0189	-0.0015	-0.0014	0.0029
	Final Term		0.0072	-0.0006	-0.0005	0.0011
Presented concepts clearly	Sessional	5479	0.0798**	-0.0080**	-0.0056**	0.0136**
	Mid Term		0.0353**	-0.0035**	-0.0025**	0.0060**
	Final Term		0.0286*	-0.0029*	-0.0020*	0.0049*
Communicated effectively	Sessional	5473	0.0700**	-0.0078**	-0.0056**	0.0134**
	Mid Term		0.0253*	-0.0028*	-0.0020*	0.0048*
	Final Term		0.0351**	-0.0039**	-0.0028**	0.0067**

Note: Standard errors are in parentheses. ** p<.01, * p<.05. The model is estimated using ordered logit model estimation techniques, with additional control variables such as class size as well as teacher designation, degree program, and semester fixed effects. Obs. and Coef. stand for Observations and Coefficients respectively. Authors' own calculation based on QEC surveys and assessment data at the university.

The results generally support H_1, indicating a significant positive correlation between Students' Evaluations of Teaching scores and their overall academic performance. Regarding H_2, the association between SET scores and performance in Sessional exams is consistently strong and positive. For midterm exams, the association is somewhat ambiguous, while for final term exams, the association is weaker or non-existent relative to the other components. These findings

provide valuable insights into the nuanced relationship between students' perceptions and their academic performance across different evaluation components.

Instructor Interaction with and Support to Students

In Panel A, Table 7, the ordered logit regression model results for the instructor's handling of students' problems reveal a statistically significant positive association with overall Students' Evaluation of Teaching (SET) scores (Coefficient = 0.0394, $p < 0.01$). The marginal effects analysis further elucidates that a one-unit increase in total class marks percentage corresponds to a 0.0047-point decrease in the probability of students disagreeing with the instructor's effectiveness, a 0.0042-point decrease in neutrality, and a 0.0088-point increase in agreement. Similarly, for the aspect of respecting and encouraging class participation, the coefficient is 0.0532 ($p < 0.01$), signifying a positive relationship with SET scores. The marginal effects illustrate that a higher class marks percentage is associated with a decrease in disagreement by 0.0043 points, a decrease in neutrality by 0.0034 points, and an increase in agreement by 0.0078 points. Furthermore, for the availability for consultation, the coefficient is 0.0271 ($p < 0.01$), and the marginal effects suggest a decrease in disagreement by 0.0022 points, a decrease in neutrality by 0.0023 points, and an increase in agreement by 0.0045 points for every one-unit increase in class marks percentage.

Table 7

Ordered Logit Regression Model Results: Student Interaction and Support

				Marginal effects		
Outcome Variable: Instructor	**Independent Variable:** Class Marks (in %)	**Obs.**	**Coef.**	**Disagree**	**Neutral**	**Agree**
Panel A						
Dealt with students' problems	Total	5454	0.0394**	-0.0047**	-0.0042**	0.0088**
Respected and encouraged class participation		5548	0.0532**	-0.0043**	-0.0034**	0.0078**
Was available for consultation		5557	0.0271**	-0.0022**	-0.0023**	0.0045**
Panel B						
Dealt with students' problems	Sessional	5454	0.0903**	-0.0106**	-0.0096**	0.0202**
	Mid Term		0.01747	-0.0021	-0.0018	0.0039
	Final Term		0.0350**	-0.0041**	-0.0037**	0.0078**
Respected and encouraged class participation	Sessional	5548	0.1211**	-0.0097**	-0.0078**	0.0175**
	Mid Term		0.0345*	-0.0028*	-0.0022*	0.0050*
	Final Term		0.0394**	-0.0032**	-0.0025**	0.0057**
Was available for consultation	Sessional	5557	0.0979**	-0.0079**	-0.0082**	0.0162**
	Mid Term		0.0002	-0.0000	0.0000	0.0000
	Final Term		0.0142	-0.0011	-0.0012	0.0023

Note. Standard errors are in parentheses. ** p<.01, * p<.05. The model is estimated using ordered logit model estimation techniques, with additional control variables such as class size as well as teacher designation, degree program, and semester fixed effects. Obs. and Coef. stand for Observations and Coefficients respectively. Authors' own calculation based on QEC surveys and assessment data at the university.

Moving to Panel B, where the analysis dissects total class marks into components, the sessional marks exhibit the highest coefficients across all three aspects: dealing with students' problems (0.0903, $p < 0.01$), respecting and encouraging class participation (0.1211, $p < 0.01$), and being available for consultation (0.0979, $p < 0.01$). These results align with

Wahid, *et al.* (2023), who found that students' perceptions of instructor-student interactions, particularly in sessional assessments, influence SET scores more than content expertise. For the handling of students' problems, the marginal effects indicate that a one-unit increase in sessional marks percentage is associated with a 0.0106-point decrease in disagreement, a 0.0096-point decrease in neutrality, and a 0.0202-point increase in agreement. Similar patterns are observed for the other aspects, emphasizing the substantial impact of sessional marks on SET scores. Midterm exam marks also exhibit significant coefficients, but their impact is generally smaller compared to sessional marks. For final term exam marks, the coefficients are generally smaller, indicating a weaker association with SET scores compared to sessional and midterm components.

In conclusion, these findings provide support for H1, revealing a positive correlation between SET scores and overall academic performance. Moreover, H2 is partially confirmed, as sessional exams play a crucial role in shaping students' perceptions of instructors, while the impact of midterm and final term exams is more nuanced and variable across different aspects of teaching evaluation.

Instructor Classroom Management and Fairness in Assessment

In Panel A of the ordered logit regression model (Table 8), the results indicate that students' overall academic performance, as measured by total marks in percentage, significantly and positively correlates with their evaluation of teaching (i.e. SET scores) for both items: "instructor maintained a conducive learning environment" (Coefficient = 0.0420, $p < 0.01$) and "instructor was fair in examinations" (Coefficient = 0.0394, $p < 0.01$). These coefficients suggest that an increase in total marks is associated with a higher likelihood of students agreeing with these statements. This finding mirrors the cross-classified framework suggested by Huang & Cai (2023), which shows that different components of assessment have distinct impacts on SET scores. Specifically, for the first item, a one-unit increase in total marks is associated with a 0.0420 increase in the log-odds of agreeing that the instructor maintained a conducive learning environment. Similarly, for the second item, a one-unit increase in total marks is associated with a 0.0394 increase in the log-odds of agreeing that the instructor was fair in examinations.

Moving to Panel B, where we dissect the total marks into its components, we observe that the sessional marks have the highest coefficient for both items. For "instructor maintained a conducive learning environment," the coefficient for sessional marks is 0.0927 ($p < 0.01$), indicating that an increase in sessional marks is associated with a higher likelihood of agreement on this statement. The coefficients for midterm and final term exam marks are comparatively smaller, with midterm being statistically insignificant. Similarly, for the item "instructor was fair in examinations," the sessional marks again have the highest coefficient (0.0686, $p < 0.01$), suggesting a stronger association between sessional marks and the perception of fairness in examinations. The coefficients for midterm and final term exam marks are smaller, with the midterm being statistically significant.

In terms of marginal effects, for both items, the positive marginal effects for sessional marks indicate a higher probability of students agreeing with the statements when their sessional marks increase. Conversely, the coefficients for midterm and final term exam marks have smaller marginal effects, implying a weaker influence on students' agreement with the statements.

In conclusion, these findings support H1, indicating a positive correlation between SET scores and overall academic performance. Additionally, H2 is partially supported, as sessional exams exhibit a stronger association with SET scores compared to midterm and final exams. The identification of sessional marks as the component with the highest coefficient highlights its significant role in influencing Students' Evaluation of Teaching.

Table 8

Ordered Logit Regression Model Results: Classroom Management and Fairness

Outcome Variable: Instructor	**Independent Variable:** Class Marks (in %)	**Obs.**	**Coef.**	**Marginal effects**		
				Disagree	**Neutral**	**Agree**
Panel A						
Maintained a conducive learning environment	Total	5518	0.0420**	-0.0038**	-0.0035**	0.0073**
Was fair in examinations		5433	0.0394**	-0.0049**	-0.0034**	0.0083**
Panel B						
Maintained a conducive learning environment	Sessional	5518	0.0927**	-0.0083**	-0.0077**	0.0160**
	Mid Term		0.0237	-0.0021	-0.0020	0.0041
	Final Term		0.0326**	-0.0029**	-0.0027**	0.0057**
Was fair in examinations	Sessional	5433	0.0686**	-0.0085**	0.0060**	0.0144**
	Mid Term		0.0247*	-0.0030*	-0.0021*	0.0052*
	Final Term		0.0383	-0.0047**	-0.0033**	0.0081**

6. Discussion

This study represents a pioneering effort to delve into the intricate dynamics of student-teacher relations and their potential influence on academic assessment, specifically focusing on the pervasive issue of grade inflation. According to Park & Cho (2022), grade inflation distorts student expectations and, consequently, SET outcomes, which this study confirmed in the context of sessional marks. To the best of our knowledge, this research is the first in this field of study that identifies the areas of assessment, systematically, that is susceptible to manipulation by both parties (i.e. teachers and students). The study investigated the relationship between Teachers' Assessment of Learning (TAL) and Students' Evaluation of Teaching (SET) scores. The findings of this study revealed a multifaceted web of factors that may influence the academic performance and shape students' perceptions on the quality of teaching. Some of the elements including the timing of assessments (i.e. exams), survey administration, disclosure of the assessment and evaluation scores to both students and teachers, and the components in the overall assessment emerged as critical factors that potentially cause manipulations by both parties (i.e. teachers and students).

Hutchinson et al. (2024) discuss the mental health and professional confidence impacts of anonymous SET feedback, which was not directly examined in this study but could offer a future direction. Particularly, the pressures on instructors to score well on SETs may lead to grade inflation, as suggested by Ikram & Kenayathulla (2023). The findings of this study are aligned with those of the work of Berezvai *et al.* (2021) that confirm that there exist an association between grading leniency and SET scores. However, our research adds this understanding further by interpreting that this effect is pronounced in sessional marks particularly that is the 30% of the total assessment. The vulnerability of sessional assessments to manipulation, attributed to the disposal of records at the semester's end, underscores the urgency of re-evaluating the weight assigned to this assessment component.

However, Graf (2024) contends that SET scores remain valid for assessing student satisfaction with instructors, even if not ideal for measuring teaching quality, suggesting that their use in career decisions can still be justified. This perspective contrasts with Uttl (2024), who argues that SET scores should not be used for high-stakes career decisions. Our study supports Uttl's caution regarding SET reliance in evaluations, especially given the potential for grade inflation tied to

sessional marks.

Contrary to established expectations and institutional review requirements, our research exposes a discernible difference in the rigor of assessment design across sessional, midterm, and final term exams. While sessional assessments are more susceptible to manipulation, the detailed nature of midterm and final assessments, integral to institutional records, aligns more closely with course objectives. This reinforces the hypothesis proposed by Clayson (2022) that SET scores predominantly measure students' likeability of their teacher rather than the effectiveness of teaching.

A distinctive contribution of this study lies in its exploration of the correlation between specific teacher characteristics and assessment components. Surprisingly, teacher subject knowledge demonstrated a weak correlation with student assessment scores, whereas personal characteristics, such as addressing students' problems and encouraging class participation, exhibited a positive correlation, particularly in sessional assessments. This finding is consistent with Wahid, Ullah, & Ranjha (2023), who observed that student perceptions in SETs often hinge on instructor-student interactions rather than content expertise. This suggests that, in sessional assessments, students' evaluations are more influenced by the teacher-student relationship than by the subject matter expertise.

Conducted at one public sector university Pakistan, this research introduces a novel approach to identify areas of assessment prone to grade inflation, emphasizing the importance of institutional policies governing assessment methods, timing, result disclosure, and accountability mechanisms. Hutchinson et al. (2024) provide additional insight into how SET systems, when not carefully managed, can negatively affect faculty mental health and professional confidence, particularly when feedback becomes anonymous and non-constructive. As Ikram & Kenayathulla (2023) suggest, relying on SETs alone can mislead decision-makers, potentially prompting faculty to inflate grades to achieve better SET results. With these finding the study cautions to the policy makers that they should avoid to rely solely on the raw SET scores which may lead both faculty members and administrators astray in their decisions. For example, the desire to get a higher SET score may drive a faculty member to inflate the sessional marks of students rather than focusing on the improvement of their teaching practices genuinely. Similarly, on the other hand, students could manipulate their teacher to get good sessional marks in return to scoring their teacher a higher SET scores. Thus, the use of student evaluations as a sole tool to assess the effectiveness and quality of teaching may not be appropriate (Zabaleta, 2007).

The findings of this study are consistent with prior studies (e.g., Berezvai et al., 2021; Stark & Freishtat, 2014), together all suggest a triangulated evaluation approach that combines SET scores with other complementary tools such as teaching portfolios, peer reviews, classroom observations, etc. Such a multidimensional framework is essential for informed policy-making and effective assessment of teaching quality and learning outcomes in HEIs in Pakistan. The findings of this study also challenge the validity of SET as a standalone metric and highlights its role in fostering an environment of grade inflation, work deflation and unhealthy competition among faculty. This study also endorses more robust and balanced evaluation methods (e.g. Uttl, 2024) to uphold academic standards.

The key theoretical contribution of this study is to build a matrix named as *'Power Matrix of Manipulation in SET and TAL: Teachers vs Students'*. This matrix introduces a novel framework to understand and categorize the ways in which this manipulation can be happened within the framework of SET and TAL. In fact, it is based on the premise that both students and teachers can influence each other to varying extents; particularly across different types of students' academic assessments or examination results. Such manipulation can potentially distort SET scores and

contribute to grade inflation which eventually undermines the validity and reliability of these measures as indicators for teaching quality and student learning.

The Power Matrix Framework reveals that certain components of TAL such as sessional (i.e. formative) assessments such as quizzes, assignments and presentations are highly vulnerable to this types of reciprocal manipulation. Where instructors with substantial discretion over sessional grading may inflate marks for securing a favourable SET scores. While on the other hand, students may use SET responses to influence their teachers on securing higher grades. This mutual influence undermines the reliability of SET surveys as a standalone measure for teaching effectiveness in HEI in Pakistan. Our framework also accounts for the role of institutional specific policies such as timing of conduct of exams or SET and dissemination of their scores that shape these dynamics. At the University of Turbat, SET surveys are conducted two weeks before the final exams and disclosed post-grading that may potentially create a power asymmetry which affects both assessment practices and evaluation outcomes.

In addition to the theoretical contribution, the matrix is also empirically validated by analysing the relationships between SET scores and the various components of TAL such as sessional (i.e. formative assessments), midterm and final exams (i.e. summative assessments). The matrix is divided into Quadrants that represent different power dynamics.

The Quadrant I (i.e. the final exams) reflects the area of minimal manipulation, as high-stakes of summative assessments by the teachers are rigorously monitored and aligned with the Intended Learning Outcomes (ILOs). Here the lack of a significant correlation between SET scores and final marks supports our argument that final term exam is a reliable indicator of teaching effectiveness. The Quadrants II/III (i.e. midterm exams) show an indeterminate relationship which is influenced by the timing of result disclosure. In contrast, the Quadrant IV (i.e. sessional assessments) reveals a strong and positive correlation between SET scores and formative marks which highlights an area of high susceptibility to reciprocal manipulation: where students reward leniency with favorable evaluations and teachers inflate marks in exams to secure higher SET scores. These findings validate our Power Matrix as a robust theoretical framework to analyze the interplay between evaluation of teaching and assessment of learning.

7. Conclusion

This study examined the interplay between Students' Evaluation of Teaching (SET) and Teachers' Assessment of Learning (TAL) in a public sector university in Pakistan. It identified sessional assessments as especially prone to reciprocal manipulation. The findings, supported by the Power Matrix framework, emphasize the risks of using SET scores in isolation for academic decisions and advocate for a more holistic evaluation approach incorporating multiple tools such as peer reviews and teaching portfolios.

This study contributes in the growing academic literature of the field of educational assessment and evaluation in the following ways: first, the Power Matrix of Manipulation in SET and TAL helps us in framing the reciprocal relationship between students and teachers in terms of compromising the academic integrity, especially in area of the sessional (formative) component of assessments. Second, the research findings validate that the sessional component of assessments is more vulnerable to manipulation which emphasizes the need for reform in integrating formative assessments into final evaluations. Third, the study also underscores the role of institutional policies—such as survey timing and score disclosure—in shaping SET–TAL dynamics further calls for policy alignment to ensure fair evaluation in HEIs. Last, unlike the traditional aggregate analysis, the study also develops a thematic categorization of SET items that offers a deeper understanding of how specific teaching attributes (e.g., punctuality, engagement, etc.) can influence SET scores.

The results of the ordered logit regression models revealed that there exists a positive correlation between SET scores and overall academic performance of students, with sessional marks is particularly highlighted. This finding support the hypothesis that students' evaluation of their teachers is positively influenced by their performance in sessional exams. The study also uncovered a huge disparity in the rigor of assessment design across sessional, midterm, and final term exams. Sessional assessment (i.e. a formative assessment in nature) was identified as a component of overall assessment that is more susceptible to manipulation by both stakeholders (i.e. teachers and students) that raises a very serious concerns about the potential distortion of evaluation scores in the existing system of higher education institutions in the country. These findings not only question the reliability of SET scores as indicators of teaching effectiveness but also point to other broader systemic issues within the assessment practices of HEIs that calls for critical reflection on current evaluation methodologies.

Based on these findings, our study advocates to use a triangulation approach for the evaluation and monitoring of quality in higher education that combines the SET scores with other evaluation tools such as teaching portfolios, classroom observation, peer reviews, document analysis, etc. By using such a multifaceted approach can offer a more holistic understanding about effectiveness of teaching and student learning outcomes that mitigates the risks associated with the potential biases and manipulations—identified in this study. In conclusion, the research contributes nuanced insights to the literature on the discussion of the relationship between students' perception of teaching and their academic performance that emphasizes the need to develop a comprehensive and context-specific approach to educational assessment and evaluation in higher education institutions. The introduction of the 'Power Matrix of Manipulation in SET and TAL' provides a novel theoretical framework to better understand the reciprocal dynamics at play that offers HEIs a valuable tool to address the complexities of evaluation and to safeguard the integrity of both teaching evaluations and learning assessments.

DECLARATION STATEMENTS

Conflict of Interest

The authors declare no actual or perceived conflicts of interest. They also confirm that no external funding was received for this study, beyond the allocation of academic time at their respective university.

Data Availability Statement

The data used in this study will be provided by the corresponding author on request.

8. References

Andrade, E. d. C., & Rocha, B. d. P. (2012). Factors affecting the student evaluation of teaching scores: Evidence from panel data estimation. *Estudos Econômicos (São Paulo)*, *42*(1), 129-150. https://doi.org/10.1590/S0101-41612012000100005

Batool, Z., & Qureshi, R. H. (2007). Quality assurance manual for higher education in Pakistan. Higher Education Commission, Pakistan.

Berezvai, Z., Lukáts, G. D., & Molontay, R. (2021). Can professors buy better evaluation with lenient grading? The effect of grade inflation on student evaluation of teaching. *Assessment & Evaluation in Higher Education, 46*(5), 793-808. https://doi.org/10.1080/02602938.2020.1821866

Binderkrantz, A. S., Bisgaard, M., & Lassesen, B. (2022). Contradicting findings of gender bias in teaching evaluations: evidence from two experiments in Denmark. *Assessment & Evaluation in Higher Education, 47*(8), 1345-1357. https://doi.org/10.1080/02602938.2022.2048355

Cannon, E., & Cipriani, G. P. (2022). Quantifying halo effects in students' evaluation of teaching. *Assessment & Evaluation in Higher Education, 47*(1), 1-14. https://doi.org/10.1080/02602938.2021.1888868

Cho, D., & Cho, J. (2017). Does more accurate knowledge of course grade impact teaching evaluation? *Education Finance and Policy, 12*(2), 224-240. https://doi.org/10.1162/EDFP_a_00197

Clayson, D. (2022). The student evaluation of teaching and likability: what the evaluations actually measure. *Assessment & Evaluation in Higher Education, 47*(2), 313-326. https://doi.org/10.1080/02602938.2021.1909702

Ewing, A. M. (2012). Estimating the impact of relative expected grade on student evaluations of teachers. *Economics of Education Review, 31*(1), 141-154. https://doi.org/10.1016/j.econedurev.2011.10.002

Fullerton, A. S. (2009). A conceptual framework for ordered logistic regression models. *Sociological Methods & Research, 38*(2), 306-347. https://doi.org/10.1177/0049124109346162

Gatwiri, K., Anderson, L., & Townsend-Cross, M. (2024). 'Teaching shouldn't feel like a combat sport': how teaching evaluations are weaponised against minoritised academics. *Race Ethnicity and Education, 27*(2), 139-155. https://doi.org/10.1080/13613324.2021.1890560

Graf, P. (2024). Making sense of today's use of student evaluations of teaching (SET). *Human Arenas, 7*(2), 446-450. https://doi.org/10.1007/s42087-023-00377-z

Harrell, F. E. (2001). Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis (Vol. 608): Springer. https://doi.org/10.1007/978-1-4757-3462-1

He, J., Zheng, X., Liu, M., Du, Y., Liu, G., Cui, J., & Su, Y. (2022). Reciprocity in college teaching: a big data study based on online student evaluation of 919,750 professors. *Assessment & Evaluation in Higher Education, 47*(8), 1401-1415. https://doi.org/10.1080/02602938.2022.2067980

Heffernan, T. (2022). 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. *Nature Reviews Psychology*. https://doi.org/10.1038/s44159-025-00444-y

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression. John Wiley & Sons. https://doi.org/10.1002/9781118548387

Huang, S., & Cai, L. (2023). Cross-Classified Item Response Theory modeling with an application to student evaluation of teaching. *Journal of Educational and Behavioral Statistics, 49*(3), 311-341. https://doi.org/10.3102/10769986231193351

Hutchinson, M., Coutts, R., Massey, D., Nasrawi, D., Fielden, J., Lee, M., & Lakeman, R. (2024). Student evaluation of teaching: reactions of Australian academics to anonymous non-constructive student commentary. *Assessment & Evaluation in Higher Education, 49*(2), 154-164. https://doi.org/10.1080/02602938.2023.2195598

Ikram, M., & Kenayathulla, H. B. (2023). Education quality and student satisfaction nexus using instructional material, support, classroom facilities, equipment and growth: Higher education perspective of Pakistan. *Frontiers in Education*. 8:1140971. https://doi.org/10.3389/feduc.2023.1140971

Jin, J. C. (2019). Student evaluation of teaching in higher education: Evidence from Hong Kong. *International Journal of Higher Education, 8*(5), 95-109. https://doi.org/10.5430/ijhe.v8n5p95

Johnson, V. E. (2006). Grade inflation: A crisis in college education: Springer Science & Business Media. https://doi.org/10.1007/b97309

Krautmann, A. C., & Sander, W. (1999). Grades and student evaluations of teachers. *Economics of Education Review, 18*(1), 59-63. https://doi.org/10.1016/S0272-7757(98)00004-1

Liu, Y., Visone, J., Mongillo, M. B., & Lisi, P. (2019). What matters to teachers if evaluation is meant to help them improve? *Studies in Educational Evaluation, 61*(June 2019), 41-54. https://doi.org/10.1016/j.stueduc.2019.01.006

Michela, J. L. (2023). Toward understanding and quantifying halo in students' evaluation of teaching. *Assessment & Evaluation in Higher Education, 48*(4), 438-450. https://doi.org/10.1080/02602938.2022.2086965

Pineda, P., & Steinhardt, I. (2023). The Debate on student evaluations of teaching: global convergence confronts higher education traditions. *Teaching in Higher Education, 28*(4), 859-879. https://doi.org/10.1080/13562517.2020.1863351

Seldin, P. (1993). The use and abuse of student ratings of professors. *The Chronicle of Higher Education, 39*(46), A40.

Sigurdardottir, M. S., Rafnsdottir, G. L., Jónsdóttir, A. H., & Kristofersson, D. M. (2023). Student evaluation of teaching: gender bias in a country at the forefront of gender equality. *Higher Education Research & Development, 42*(4), 954-967. https://doi.org/10.1080/07294360.2022.2087604

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research, 83*(4), 598-642. https://doi.org/10.3102/0034654313496

Stark, P. B., & Freishtat, R. L. (2014). An evaluation of course evaluations. *ScienceOpen Research, 0*(0) 1-7. https://doi.org/10.14293/S2199-1006.1.SOR-EDU.AOFRQA.v1

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

Sullivan, D., Lakeman, R., Massey, D., Nasrawi, D., Tower, M., & Lee, M. (2024). Student motivations, perceptions and opinions of participating in student evaluation of teaching surveys: A scoping review. *Assessment & Evaluation in Higher Education, 49*(2), 178-189. https://doi.org/10.1080/02602938.2023.2199486

Uttl, B. (2024). Student evaluation of teaching (SET): Why the emperor has no clothes and what we should do about it. *Human Arenas, 7*(2), 403-437. https://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. https://doi.org/10.1016/j.stueduc.2016.08.007

Wahid, A., Ullah, R., & Ranjha, A. N. (2023). Assessing the effectiveness of students' evaluation of teachers: A qualitative analysis of Pakistan's higher education. *Meritorious Journal of Social Sciences and Management, 6*(01), 9-19.

https://journal.mgp.org.pk/index.php/mjssm/article/view/178

Weinberg, B. A., Hashimoto, M., & Fleisher, B. M. (2009). Evaluating teaching in higher education. *The Journal of Economic Education, 40*(3), 227-261. https://doi.org/10.3200/JECE.40.3.227-261

Wright, S. L., & Jenkins-Guarnieri, M. A. (2012). Student evaluations of teaching: Combining the meta-analyses and demonstrating further evidence for effective use. *Assessment & Evaluation in Higher Education, 37*(6), 683-699. https://doi.org/10.1080/02602938.2011.563279

Zabaleta, F. (2007). The use and misuse of student evaluations of teaching. *Teaching in Higher Education, 12*(1), 55-76. https://doi.org/10.1080/13562510601102131