11th International Conference on Higher Education Advances (HEAd'25)
Universitat Politècnica de València, València, 2025
DOI: https://doi.org/10.4995/HEAd25.2025.20055

Do Different Rating Scales in Student Evaluations Bias Perceptions of Teaching Quality?

Zana Demiri, Michael Burkert
Department of Management, University of Fribourg, Switzerland.

How to cite: Demiri, Z.; Burkert M. (2025). Do Different Rating Scales in Student Evaluations Bias Perceptions of Teaching Quality? In: 11th International Conference on Higher Education Advances (HEAd'25). Valencia, 17-20 June 2025. https://doi.org/10.4995/HEAd25.2025.20055

Abstract

This study examines how different rating scales in Student Evaluation of Teaching (SET) influence perceived teaching quality and decision-making. In Phase 1, SET data was collected from five university courses, with students randomly assigned to evaluate their instructor using one of two different rating scales per course. In Phase 2, 3,719 online participants completed one of two decision-making tasks based on SET results from Phase 1. Participants were randomly assigned to select a university for their studies or hire a candidate, with ratings derived from varying rating scales. A linear regression analysis shows that differences in standardized scores derived from the usage of different rating scales significantly influence selection outcomes. Additionally, a thematic analysis of open-ended responses reveals that participants relied on percentage-based comparisons, but their choices were adjusted when the distribution of responses was provided. These findings uncover the bias introduced by different rating scales in SET.

Keywords: *SET, student evaluation, rating scales, bias, higher education*

1. Introduction

Management control systems (MCSs) guide organizations in aligning employee behavior with strategic goals (Merchant & Van der Stede, 2007). While extensively studied in corporate settings, their application in higher education remains underexplored. With increasing pressure on universities to ensure quality teaching, Student Evaluations of Teaching (SET) have emerged as a dominant evaluation tool, adopted by 94% of institutions in Germany, Austria, and Switzerland (Guenther & Schmidt, 2015). Despite their widespread use, SETs face criticism for biases and methodological limitations (Boring et al., 2016; Stroebe, 2020). Prior research highlights how rating scales influence psychometric properties such as reliability and validity (Cox, 1980; Schmitt & Stults, 1986). However, little attention has been given to how different response scales alter perceived teaching quality. This study addresses this gap by analyzing how scale variations affect SET results and subsequent decision-making. To investigate, we collected

SET data from courses at a Western European university, applying different response scales. We then conducted a survey where participants compared evaluation scores presented in hiring and university selection scenarios. Our goal was to understand how participants interpret and compare SET scores derived from different scales. Notably, we intentionally held teaching quality constant by design, altering only the rating scale used for evaluation.

The paper proceeds as follows: Section 2 details our methodology, Section 3 presents results, and Section 4 concludes with implications.

2. Methodology

We conducted a two-phase approach: in Phase 1, we collected real SET data using various rating scales; for Phase 2, we conducted an experimental survey to analyze how participants interpret and compare these SET results in two different decision contexts.

2.1. Phase 1

To obtain real-world SET data, we evaluated five different courses at a Western European university. Each evaluation questionnaire contained the same 19 standardized questions, and students were randomly assigned one of two rating scales tested within a single course. This setup ensured that we could directly compare evaluation results for the same instructor, in the same course, at the same time – isolating the effect of the rating scale itself. Across the five courses, students were exposed to five different scale combinations: 4-point (n=17/34) vs. 7-point (n=17/34), 4-point (n=27/49) vs. 7-point (n=22/49), 4-point (n=22/44) vs. 8-point (n=22/44), 5-point (n=10/27) vs. 7-point (n=17/27), and 4-point (n=13/22) vs. 6-point (n=9/22), resulting in a total of 176 completed evaluations.

2.2. Phase 2

Using this data, we designed an experimental online survey with two decision-making scenarios to test how participants interpret the evaluation scores from Phase 1.

In the first scenario (Figure 1), participants imagined themselves as students choosing a university for their master's program. They were presented with two options, University X and University Y, and informed that both universities were comparable in quality, apart from the alleged evaluations, which entailed the master program evaluation within the two universities. These ratings, allegedly from former students, were taken from our SET dataset in Phase 1, where we assigned two different average scores from the same instructor’s evaluations—each derived from a different rating scale—to the two universities. Participants had to decide which university they would prefer, based solely on these ratings.

In the second scenario (Figure 2), participants imagined themselves as hiring committee members selecting a professor. Two candidates, Candidate A and Candidate B, were described as having identical research backgrounds, with the only differentiating factor being their teaching evaluations. As in the first scenario, we presented the teaching performance of both candidates by extracting averages from our SET dataset in Phase 1. Specifically, each candidate was assigned one of the two averages derived from the evaluation of the same instructor, who was assessed in Phase 1 using two different rating scales. In this scenario, participants were also shown the distribution of scores behind each average. This allowed us to assess how access to additional data influenced decision-making.

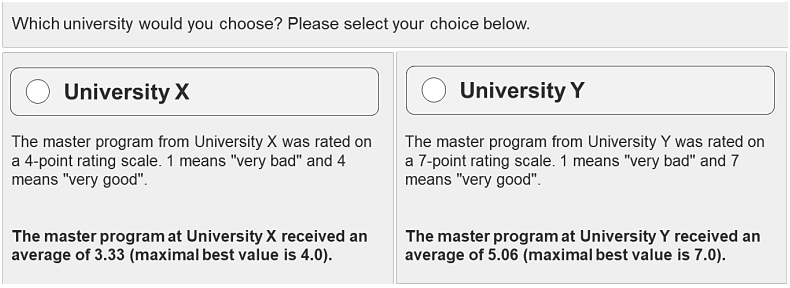

Figure 1. *Decision-making scenario: Selection between University X and University Y based on ratings from different scales. The figure displays the exact interface shown to study participants when they were asked to choose between two universities based solely on master program evaluation scores derived from different rating scales. Source: Authors' Own Study (2024).*

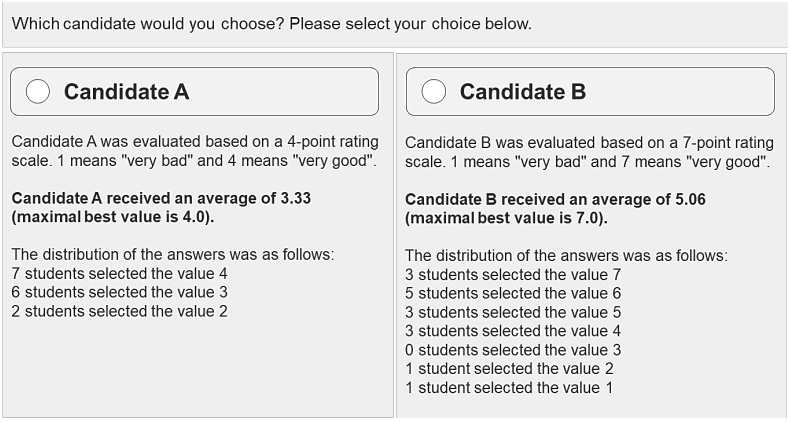

Figure 2. *Decision-making scenario: Selection between Candidate A and Candidate B based on ratings from different scales, including the distribution of responses. The figure displays the exact interface shown to study participants when they were asked to choose between two candidates based solely on teaching evaluation scores derived from different rating scales. Source: Authors' Own Study (2024).*

The survey was conducted via Prolific, a UK-based research platform, with 3,719 participants from Europe. Each participant received a small financial incentive, and the median completion time was 3 to 4 minutes.

To analyze responses, we applied both quantitative and qualitative methods. First, SET scores were standardized into percentages for comparability, and a linear regression model examined the relationship between score differences and voting outcomes. Second, a thematic analysis of 3,719 open-ended responses was conducted following Guest et al.'s (2014) framework to identify key reasoning patterns. This approach enables us to assess how small changes in rating scales impact both SET results and the decisions made based on those results. By combining empirical SET data with a survey experiment, we provide a comprehensive examination of rating scale effects in higher education assessments.

3. Results

3.1. Impact of Rating Scales on SET Scores (Phase 1)

To compare the SET evaluations across different rating scales acquired in Phase 1, we standardized the averages using the following formula, adapted from Preston & Colman (2000):

$$Standardized\ Score = \left\{ \frac{Rating-1}{Max\ Scale\ Value-1} \right\} x\ 100 \quad (1)$$

This transformation converts all SET scores into a 0 - 100% scale, allowing for direct comparisons. For example, one instructor received an average score of 3.80/5, which - when standardized using Formula (1) - became 70.00%. The same instructor, when evaluated using a 7-point scale for the same question, received an average of 6.71/7, which standardized to 95.86%.

The analysis of 19 score comparisons between different rating scales for the same course and instructor revealed notable variations in standardized scores. When applying the standardization formula, we observed differences of up to 25 percentage points. The mean absolute difference across all rating scale comparisons was 7.93 percentage points.

These findings underscore that rating scale selection alone can significantly impact perceived teaching performance, which may in turn influence faculty evaluations, hiring decisions, and institutional policies.

3.2. Survey Experiment Findings (Phase 2)

A total of 3,719 participants completed the survey, making decisions in two scenarios based on SET-derived ratings from Phase 1.

In the first scenario, where participants chose a university based solely on its average student evaluation score, 95% of participants selected the university with the higher standardized percentage rating. This confirms that participants mentally convert different rating scales into a common percentage-based format before making decisions.

In the second scenario, where participants evaluated faculty candidates and were also shown the distribution of scores, 89% of participants selected the candidate with the higher percentage rating – similar to the first scenario. Therefore, the assumption that most of the votes were given to the university/candidate with the highest percentage rate holds true for both scenarios.

3.3. Regression Analysis

A linear regression model was used to examine the relationship between score differences and voting outcomes. Results, based on all data from Phase 2, show a strong positive relationship between the difference in standardized scores and the difference in votes. Specifically, a 1% increase in the standardized score difference resulted in a 4.70% increase in the vote difference ($p < 0.001$). This confirms that the comparison of evaluation results, stemming from different rating scales, systematically influence voting outcomes. Table 1 illustrates this relationship, showing a case where Candidate B received a standardized score that was 25.17% higher than Candidate A (95.17% vs. 70.00%). This difference in standardized scores translated into a 77.08% difference in votes (88.54% vs. 11.46%). The results align with the model's findings, reinforcing the conclusion that differences in evaluation results due to varying rating scales can have a measurable impact on decision-making.

Table 1. Example of How Differences in Standardized Scores Translate into Voting Outcome.
Source: Own elaboration based on Phase 2 data.

	Standardized Score (%)	Standardized Score Difference (%)	Votes	Votes (%)	Vote Difference (%)
Candidate A Score 3.80/5	70.00%	-25.17%	11	11.46%	-77.08%
Candidate B Score 6.71/7	95.17%	+25.17%	85	88.54%	+77.08%

3.4. Thematic Analysis of Open-Ended Responses

We asked all participants an open-ended question about the reasoning behind their choice. Using thematic analysis, we aimed to uncover the multiple factors influencing voting behavior in greater detail. The thematic analysis followed the framework by Guest et al. (2014) and was conducted on a dataset of open-ended responses from a total of 3,719 study participants. This analysis identifies key themes explaining how participants evaluate universities and faculty candidates based on SET scores measured on different scales, with Table 2 presenting the main factors that emerged as influencing their decisions.

In Scenario 1 (University Selection), participants were provided only with average scores. Decision-making in this scenario was driven primarily by quantitative metrics, as participants overwhelmingly preferred higher percentage ratings, top scores, and standardized scores. This suggests a strong reliance on numerical comparisons when no additional context was available. In Scenario 2 (Faculty Hiring), where participants also received score distribution data, decision-making patterns shifted. Participants placed greater emphasis on minimizing negative feedback, often favoring candidates with fewer negative ratings even when their overall average was lower. Additionally, evaluation reliability, as indicated by larger sample sizes, played also a more prominent role in their choices.

Table 2. Identified Decision-Making Themes in the Thematic Analysis. Source: Own elaboration based on responses from 3,719 study participants in Phase 2.

Theme	Description	Scenario
Higher Percentage	Preference for options with a higher percentage rate, indicating quantitative superiority.	1 & 2
Standardized Score	Adjustment of scores to a common standard to compare options across different scales.	1 & 2
Top Score	Preference for options closer to the maximum possible score on its scale.	1 & 2
Large Scale	Preference for evaluations on a larger scale, perceived as more comprehensive.	1 & 2
Distribution More Positive	Preference for options with a higher proportion of positive evaluations.	2
Distribution Less Negative	Preference for options with fewer negative evaluations, minimizing poor outcomes.	2
Sample Size	Preference for options evaluated by a larger number of participants, seen as more reliable.	2

To analyze decision patterns, we categorized universities and faculty candidates into green and red categories based on their difference in standardized scores. The green category included candidates or universities with a positive difference, meaning they received a higher standardized score compared to their counterpart. Conversely, the red category included those with a negative difference, indicating a lower standardized score. In the example from Table 1, Candidate B falls into the green category, having a higher standardized score, while Candidate A is in the red category, with a lower standardized score. This classification allowed for a structured examination of participants' choices and the reasoning behind them. Across both scenarios, the green category options were primarily chosen based on high standardized scores, whereas red category options were selected when participants considered factors such as rating scales, sample size, and negative score distribution. This finding suggests that while score-based comparisons dominate decision-making, the distribution of responses and sample size become equally important once this additional information is provided to decision-makers.

3.5. Excluded Themes

Some participant responses did not fit into the structured decision-making themes presented in Table 2. Several participants relied on intuition, making decisions based on gut feelings rather than systematic comparison. Others provided no clear reasoning, selecting an option arbitrarily without justification. Additionally, some responses were irrelevant or incoherent, making them unsuitable for analysis. While a few participants referenced "better ratings," their criteria for defining "better" were often vague, making systematic categorization difficult.

4. Conclusion

This study examined how variations in rating scales influence Student Evaluations of Teaching (SETs) and the decision-making processes that follow. The findings demonstrate that even when the instructor, course content, and student cohort remain constant, differences in rating scales can lead to notable changes in standardized scores. This suggests that the same teaching performance may be perceived quite differently depending on the scale used. Through a survey experiment, we found that participants predominantly relied on numerical averages when evaluating universities and faculty candidates, often converting scores into percentages and choosing the option with the higher value. However, when additional information such as score distributions was available, participants became more attentive to negative feedback and the reliability of evaluations based on sample size. These patterns illustrate the risks of comparing SETs across different scale formats without proper context. Institutions using SETs for hiring, promotion, or rankings should therefore exercise caution, as scale design alone can influence decision outcomes. This aligns with Rivera and Tilcsik's (2019) findings that subtle differences in rating scales, such as the number of points, can introduce bias—particularly gender-related bias—in teaching evaluations. Future research could address ways to mitigate these effects, for example, through standardized rating scales or by providing contextual explanations with evaluation results. While this study is limited by the number of courses included and the specific range of scale formats tested, it underscores the broader implications of rating scale design. As SETs continue to play a critical role in assessing teaching quality, ensuring their fair and meaningful interpretation remains essential for academic institutions and faculty alike.

References

Boring, A., Ottoboni, K., & Stark, P. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research*, *2016*, 1–11. https://doi.org/10.14293/s2199-1006.1.sor-edu.aetbzc.v1

Cox, E. P. (1980). The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research*, *17*(4), 407–422. https://doi.org/10.2307/3150495

Guenther, T. W., & Schmidt, U. (2015). Adoption and use of management controls in higher education institutions. In B. R. Martin, D. M. H. Grant, & U. Schmoch (Eds.), *Incentives and performance: Governance of research organizations* (pp. 213–236). Springer. https://doi.org/10.1007/978-3-319-09785-5

Guest, G., MacQueen, K. M., & Namey, E. E. (2012). *Applied thematic analysis*. SAGE Publications. https://doi.org/10.4135/9781483384436

Merchant, K. A., & Van der Stede, W. A. (2007). *Management control systems* (2nd ed.). Pearson Education Limited.

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. https://doi.org/10.1016/S0001-6918(99)00050-5

Rivera, L. A., & Tilcsik, A. (2019). Scaling down inequality: Rating scales, gender bias, and the architecture of evaluation. *American Sociological Review, 84*(2), 248–274. https://doi.org/10.1177/0003122419833601

Schmitt, N., & Stults, D. M. (1986). Methodology review: Analysis of multitrait-multimethod matrices. *Applied Psychological Measurement*, *10*(1), 1–22. https://doi.org/10.1177/014662168601000101

Stroebe, W. (2020). Student evaluations of teaching encourage poor teaching and contribute to grade inflation: A theoretical and empirical analysis. *Basic and Applied Social Psychology*, *42*(4), 276–294. https://doi.org/10.1080/01973533.2020.1756817

Appendix A

1	I regularly attend the course.
2	The learning objectives of the course are clear.
3	The course has a clear structure.
4	The instructor is committed to their teaching.
5	The course provides a learning-conducive atmosphere.
6	The evaluation methods (exams, assignments, etc.) are aligned with the learning objectives.
7	The course enables me to deepen the content independently.
8	The offered activities (readings, group work, excursions, etc.) support my learning.
9	I have the impression that the course prepared me well for the assessment (exam, assignment, etc.).
10	The content is presented in an understandable way.
11	The material is illustrated with examples
12	The requirements are...
13	I actively participate in the course (thinking along, preparing and reviewing, regular attendance, etc.)
14	The videos and interactive case studies helped me to better understand the material.
15	Mother tongue(s)
16	Gender
17	Degree program
18	I am taking this course as part of…
19	For me, this course is…