Decision Analytics Journal 21 (2026) 100756

Contents lists available at ScienceDirect

Decision Analytics Journal

journal homepage: www.elsevier.com/locate/dajour

An analytics-driven outranking method for bias mitigation in survey-based performance assessment

Felipe Barrera-Jimenez [a,*], Jose Luis Arroyo-Barrigüete [a], Marina Segura [b], Antonio Obregón-García [a], Carlos Álvarez-Fernández [a], Jose María Ortiz-Lozano [a]

[a] Universidad Pontificia Comillas, Santalucía Chair of Analytics for Education, Calle de Alberto Aguilera, 23, 28015, Madrid, Spain
[b] Universidad Complutense de Madrid, Campus de Somosaguas, 28223, Pozuelo de Alarcón, Madrid, Spain

ARTICLE INFO

Keywords:
Survey analytics
Multicriteria decision analysis
Outranking methods
Bias mitigation
Preference ranking method
Teaching performance assessment

ABSTRACT

Survey-based evaluations inform high-stakes ranking and performance decisions, yet scores often reflect non-instructional influences. This paper introduces PROMSET (Preference Ranking Organization Method for Student Evaluation of Teaching), a bias-mitigated outranking method that integrates robust multiple linear regression for bias identification with a comparison-dependent extension of PROMETHEE II. The method is validated on Student Evaluation of Teaching (SET) data from a Spanish university (45,357 surveys; 879 course evaluations; four knowledge areas). PROMSET incorporates a comparison-dependent activation mechanism that applies bias-calibrated indifference thresholds only when pairwise comparisons involve bias-favored alternatives. The results show that grade leniency bias and professor gender bias are no longer statistically detectable at conventional significance levels ($p > 0.05$) under PROMSET, while the discriminating power of the ranking increases substantially. Benchmarks against the traditional method and PROMETHEE II confirm that bias remains statistically significant under conventional approaches but not under PROMSET. PROMSET improves the informational quality and fairness of survey-based rankings for institutional decision-making.

1. Introduction

Survey-based evaluation systems are widely used to support ranking and performance decisions across domains. As these systems rely on aggregated subjective assessments, they are inherently exposed to systematic measurement distortions that may affect ranking outcomes. In higher education, these dynamics become particularly salient in the evaluation of university teaching.

The evaluation of university teaching is widely recognized as a critical tool for ensuring academic quality and fostering continuous improvement in educational programs [1]. In the current context of increased accountability, funding constraints, and institutional competitiveness, universities have strengthened their performance management systems to support strategic decision-making, align academic practices, and improve institutional effectiveness [2].

Within this framework, student feedback has gained significant prominence through surveys known as Student Evaluation of Teaching (SET), becoming central in assessing instructor performance [3,4]. However, concerns have also been raised regarding the reliability and stability of SET as a measure of teaching quality [5]. These evaluations are commonly employed to provide feedback to faculty and to inform decisions on promotion, tenure, remuneration, and career progression [6,7]. Stroebe [8] further argues that their use in personnel decisions can create incentives for lenient grading and contribute to grade inflation. As Chisadza et al. [9] argue, SETs "are vital for the academic career trajectories of higher education lecturers", which underscores the importance of designing evaluation systems free from bias to ensure fairness, transparency, and effectiveness in academic human resource management.

In addition to its central role in academic human resource management, SETs also play an important role in supporting the broader strategic goals of universities and schools, most notably the pursuit of international accreditations. A key requirement across these accreditation frameworks is the evaluation of teaching effectiveness. To meet this standard, institutions often rely on SET, either alongside other evaluation tools or, in many cases, as the sole assessment method. As Siddiqui and Lento [10] note, even institutions that adopt multiple indicators of teaching effectiveness tend to focus on SET.

SET surveys comprise multiple items designed to assess various aspects of teaching, including clarity of instruction, course organization, faculty engagement, and learning outcomes. This multidimensional structure allows institutions to capture distinct facets of instructional performance, resulting in a more nuanced and comprehensive view of teaching quality [11]. Moreover, this same structure supports the

* Corresponding author.
E-mail addresses: fbarrera@comillas.edu (F. Barrera-Jimenez), jlarroyo@comillas.edu (J.L. Arroyo-Barrigüete), marina.segura@ucm.es (M. Segura), aobregon@comillas.edu (A. Obregón-García), calvarez@comillas.edu (C. Álvarez-Fernández), jmortiz@comillas.edu (J.M. Ortiz-Lozano).

https://doi.org/10.1016/j.dajour.2026.100756
Received 9 July 2026; Received in revised form 29 September 2026; Accepted 8 October 2026
Available online 13 October 2026
2772-6622/© 2026 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

use of Multiple Criteria Decision Making (MCDM) approaches. MCDM methods provide the Decision Maker (DM) with a structured framework to evaluate multiple, often conflicting, criteria in complex decision-making scenarios [12]. Each item in SET serves as a specific criterion that can be weighted and synthesized into a coherent global judgment aligned with institutional priorities [1,13].

While the multidimensional structure of SET could potentially offer a framework for evaluating teaching quality, its effectiveness depends significantly on addressing various methodological concerns. Despite their widespread use in higher education institutions, SETs face criticism due to certain limitations that may affect their validity as fully objective indicators [4,5]. One of the most significant limitations involves numerous non-instructional factors that generate systematic biases in student evaluations, as extensively documented in the literature. Among these, biases related to the disciplinary area of the course [14], Grade Leniency Bias (GLB) [15,16], and instructor-related personal characteristics, such as Gender Bias (GB) [17,18], are particularly notable. This broader pattern of bias in SETs has also been documented in relation to sexism, racism, and prejudice [19]. For instance, a one-point increase in the average course grade can add up to 0.2–0.4 units to the overall professor evaluation score [20]. When such biases remain unaddressed, there is a risk that decisions based on SET may be inappropriately guided and misaligned with institutional objectives, particularly in the context of academic human resources management.

Although these biases have been extensively studied, few works have proposed effective quantitative methods to mitigate their impacts. Moreover, adequately integrating each SET item as a specific criterion within multicriteria evaluation frameworks and channeling student feedback according to institutional preferences remains an underexplored area. Therefore, a methodological gap exists, highlighting the need to articulate both the diagnosis and mitigation of biases in SET with the generation of more reliable teaching evaluations.

The first objective of this study is to develop a multicriteria outranking method for mitigating non-instructional biases in survey-based evaluations. The proposed method, PROMSET (Preference Ranking Organization Method for Student Evaluation of Teaching), integrates robust Multiple Linear Regression (MLR) for bias detection with a structural extension of PROMETHEE II to generate bias-mitigated rankings. The second objective is to empirically validate PROMSET using a large-scale dataset of SET responses from a Spanish university, assessing its ability to mitigate two prominent non-instructional biases, GLB and GB, modeled as continuous and categorical variables, respectively.

This research provides important contributions. First, it presents a structural extension of PROMETHEE II that introduces dynamic, bias-calibrated thresholds within the preference functions through a comparison-dependent activation mechanism. This extension is jointly integrated with MLR for systematic bias identification, embedding econometric detection directly into the multicriteria aggregation process. Second, this study proposes a methodological reconceptualization of SET as an MCDM problem. It empirically validates this framework using student responses from 879 course evaluations across four knowledge areas, demonstrating PROMSET's practical viability in mitigating two of the most widely studied non-instructional biases. Finally, it includes a sensitivity analysis framework that supports institutions in selecting threshold configurations to ensure effective bias mitigation across different evaluation contexts.

The remainder of the paper is organized as follows. Section 2 reviews statistical and MCDM approaches to SET. Section 3 details the PROMETHEE methodology. Section 4 describes PROMSET in full and presents the empirical case for validation. Section 5 reports the results, which are analyzed and discussed in Section 6. Finally, Section 7 presents the conclusions.

2. Related work

This section reviews the two streams of literature on which our proposal builds. Section 2.1 examines statistical research on non-instructional biases in SET, together with the few existing attempts to correct such biases after data collection. Section 2.2 then turns to multicriteria methods applied to faculty evaluation.

2.1. Statistical methods

As previously noted, extensive literature documents the presence of non-instructional biases in SET. Among the most relevant is GLB, where students who receive higher grades tend to give better evaluations to their instructors [8,20], even though this does not necessarily reflect actual student learning. Indeed, Uttl et al. [21] found no significant relationship between SET ratings and student learning in their meta-analysis. Another common bias is related to the subject area, with instructors teaching quantitative courses often receiving lower evaluations [14,22]. Lastly, although there is ongoing academic debate, several studies suggest a negative bias against female instructors [18,23].

Certainly, a substantial body of research has focused on identifying these biases. More than ten years ago, Spooren et al. [24] observed that "bias studies continue to play a central role in the recent literature", an assertion still accurate today. Zhao et al. [25] in their recent literature review on SET in higher education similarly note that "researchers will continue to explore the influencing factors of students' teaching evaluation". On the other hand, research has also examined how SET might be supplemented or replaced by other assessment mechanisms, largely due to these non-instructional biases. However, there remains a significant gap in the literature: few studies attempt to correct the biases within SET, rather than merely suggesting ways to complement or replace them. Even the possibility of adjusting for bias has been questioned, owing to the wide range of factors that contribute to its occurrence [26].

Experimental literature has moved beyond simply identifying biases and now explores their mitigation, which may represent a more feasible approach considering the widespread reliance on SET. One approach involves preventive interventions applied before data collection. Several field experiments have shown that including brief anti-bias reminders in SET questionnaires can alter ratings. Peterson et al. [27] conducted an experiment in the U.S. and found that including a prompt about implicit gender stereotypes increased ratings for the disadvantaged group, without affecting the other group. Similarly, Boring and Philippe [28] found that a basic normative message was ineffective, whereas providing concrete historical data about discrimination reduced the gender gap. However, the replication by Kim et al. [29] cautions about potential adverse outcomes: corrective effects may lower ratings for one or more groups. Together, these studies suggest that mitigation messages can influence SET outcomes, although the direction and magnitude of their effects depend on the wording of the message and the demographic combination of students and instructors.

Proposals aimed at mitigating these biases after data collection remain limited. Admittedly, a few studies have explored this direction, such as the recent papers by Arroyo-Barrigüete et al. [30] and Zipser and Mincieli [31]. In both cases, given that MLR, typically estimated using Ordinary Least Squares (OLS), or similar methods are commonly used to identify non-instructional biases, the authors chose to use these models as a starting point to mitigate bias by analyzing the residuals. Their aim was to offer a more equitable and statistically robust interpretation of faculty performance. The results of their proposed methodologies can help address some concerns about using SET in personnel decision-making. However, few alternative proposals exist beyond these examples. Addressing this gap is precisely the goal of this paper: to propose a more robust method that, while also based on OLS models, incorporates multicriteria outranking techniques to achieve better correction of non-instructional biases.

2.2. Multicriteria methods

MCDM methods have expanded widely across several decision-making fields, including education, due to their capacity to synthesize the complexity inherent in problems involving numerous factors and stakeholders. For instance, they have been applied to assess institutional innovative performance [32], evaluate projects involving multiple stakeholders [33], rank universities or educational programs [34,35], support course planning and academic resource allocation [36], address student selection and assignment problems [37], select and evaluate e-learning platforms [38,39], and assess learning methodologies [40].

Within this framework, university faculty evaluation constitutes a particularly prolific field for MCDM methods due to the multidimensional nature of teaching performance. The literature frequently utilizes the Analytic Hierarchy Process (AHP) to determine the relative importance of evaluation criteria and integrates fuzzy techniques to model the inherent uncertainty in teacher evaluation. For example, Chen et al. [1] combined fuzzy AHP with a fuzzy comprehensive evaluation, while Qin et al. [41] adapted Interval AHP to capture the dispersion of judgments and Do et al. [42] linked fuzzy AHP with fuzzy TOPSIS to rank faculty performance. Alongside these approaches, Popović et al. [43] incorporated conjoint analysis to reflect student preferences before applying AHP and Data Envelopment Analysis (DEA).

The proliferation of methods such as Picture Fuzzy Sets with Mixed Aggregation by COmprehensive Normalization Technique (MACONT) [44], improved TOPSIS variants [45], or plithogenic hypersoft sets combined with TOPSIS [46] confirms a trend toward modeling linguistic imprecision and subjective variability. Similarly, Radovanović et al. [47] integrated Spherical Fuzzy AHP and Grey MARCOS to weight criteria and identify the best teaching option, illustrating how fuzzy frameworks can deliver robust rankings. However, these studies do not explicitly address the identification and mitigation of non-instructional biases. Furthermore, most of them require collecting linguistic judgments, detailed comparison matrices, or extensive evaluator involvement, which complicates their deployment on large datasets like those generated by SET surveys, limiting their practical institutional-scale application.

In contrast to the predominance of AHP and ideal-solution-based techniques such as TOPSIS, outranking techniques have received comparatively less attention. Mohan et al. [48] highlight the utility of PROMETHEE for ranking candidates for teaching awards, yet empirical evidence in this area remains limited. Even scarcer is the literature relying exclusively on SET data. Thanassoulis et al. [49] combined AHP and DEA on student responses to identify areas for faculty improvement, while Hadad et al. [13] used AHP to contrast the importance students and professors assign to each evaluation criterion. Bailey and Michaels [50] propose an optimization-based decision support system that balances classroom composition to reduce structural biases in future SET scores.

In summary, this review confirms the efficacy of AHP and its variants for robustly estimating criterion weights in faculty evaluation. However, adoption of outranking methods remains limited, and few studies fully exploit SET data despite its widespread use in university practices. Finally, despite advances in fuzzy modeling, a significant gap persists in directly mitigating systematic biases arising from non-instructional factors. These gaps present an opportunity to develop frameworks that leverage multicriteria rigor to integrate large-scale SET data and explicitly address non-instructional biases.

3. Methodology: PROMETHEE II

PROMETHEE II is an outranking-based multicriteria decision-making approach that evaluates preferences between alternatives through pairwise comparisons across multiple criteria. Table 1 presents the evaluation table for a set of alternatives (professors) $A = \{p_1, p_2, \dots, p_i, \dots, p_n\}$, assessed according to a set of criteria $G = \{g_1, g_2, \dots, g_j, \dots, g_m\}$. Each criterion g_j is assigned a weight $w_j \in W = \{w_1, w_2, \dots, w_j, \dots, w_m\}$, which reflects its relative importance. The objective for each criterion may be either maximization or minimization, depending on the desired performance outcome.

Table 1
Table of professor evaluation.

Professors	g_1	$\cdots$	g_j	$\cdots$	g_m
	w_1		w_j		w_m
p_1	$g_1(p_1)$	$\cdots$	$g_j(p_1)$	$\cdots$	$g_m(p_1)$
p_2	$g_1(p_2)$	$\cdots$	$g_j(p_2)$	$\cdots$	$g_m(p_2)$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
p_i	$g_1(p_i)$	$\cdots$	$g_j(p_i)$	$\cdots$	$g_m(p_i)$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
p_n	$g_1(p_n)$	$\cdots$	$g_j(p_n)$	$\cdots$	$g_m(p_n)$

Based on the evaluation table, the preference of alternative p_i over p_q is calculated for each criterion g_j. First, the performance difference is identified, as shown in Eq. (1):

$$d_j(p_i, p_q) = g_j(p_i) - g_j(p_q). \tag{1}$$

Next, this difference d_j is transformed into a preference value $P_j(p_i, p_q)$ within the range [0, 1], using a preference function F_j defined according to the evaluation requirements, as shown in Eq. (2):

$$P_j(p_i, p_q) = F_j\left[d_j(p_i, p_q)\right], \quad 0 \le P_j(p_i, p_q) \le 1. \tag{2}$$

There are different types of preference functions, and one is defined for each criterion g_j. The linear preference function (used in this article) includes a preference threshold T_j and an indifference threshold $0 < q_j < T_j$, or $q_j = 0$ (no indifference, corresponding to the V-shape function). Both V-shape and linear preference functions are particularly suitable for quantitative criteria measured on continuous scales, while other functions, such as usual and level, are generally associated with qualitative criteria or scales with few levels [51,52].

If $d_j(p_i, p_q) \le q_j$, then $P_j(p_i, p_q) = 0$. If $q_j < d_j(p_i, p_q) \le T_j$, there is a preference that increases linearly, as defined in Eq. (3). If $d_j(p_i, p_q) > T_j$, then $P_j(p_i, p_q) = 1$. For minimization criteria, the difference is defined so that larger values still indicate better performance.

$$P_j(p_i, p_q) = \frac{d_j(p_i, p_q) - q_j}{T_j - q_j}. \tag{3}$$

The next step consists of aggregating the individual preferences into a global preference index (Eq. (4)), which reflects how much alternative p_i is preferred over p_q when considering all criteria and their respective weights (weights are normalized to sum to one).

$$\pi(p_i, p_q) = \sum_{j=1}^{m} P_j(p_i, p_q) \cdot w_j. \tag{4}$$

Then, for each alternative p_i the positive and negative flows are calculated (PROMETHEE I). The positive flow $\varphi^+(p_i)$, shown in Eq. (5), measures how strongly p_i is preferred over the other alternatives. Conversely, the negative flow $\varphi^-(p_i)$, in Eq. (6), captures the extent to which p_i is less preferred than the rest:

$$\varphi^+(p_i) = \frac{1}{n-1} \sum_{p_q \in A, p_i \neq p_q} \pi(p_i, p_q), \tag{5}$$

$$\varphi^-(p_i) = \frac{1}{n-1} \sum_{p_q \in A, p_i \neq p_q} \pi(p_q, p_i). \tag{6}$$

The difference between the positive and negative flows gives the net flow $\varphi(p_i)$, as shown in Eq. (7). This value provides a complete ranking of the alternatives (PROMETHEE II). A higher net flow indicates a stronger preference for p_i, within the set A. Alternatives with $\varphi > 0$, are strongly preferred over those with $\varphi < 0$ [53].

An alternative calculation approach involves computing the net flows by criterion before applying Eq. (4). In this case, a net flow $\varphi_j(p_i)$ is obtained for each criterion g_j and the global net flow is calculated as a weighted average, as shown in Eq. (8). For further details on the PROMETHEE II method, see Brans and De Smet [54].

$$\varphi(p_i) = \varphi^+(p_i) - \varphi^-(p_i), \tag{7}$$

$$\varphi(p_i) = \sum_{j=1}^{m} \varphi_j(p_i) \cdot w_j. \tag{8}$$

4. PROMSET: A new MCDM evaluation method using [SETs] with bias mitigation

This section presents the proposed method. Section 4.1 describes the seven steps of PROMSET, from the definition of the evaluation problem to the validation of the resulting rankings. Section 4.2 then details how each of these steps was implemented in the empirical validation, based on real SET data from a Spanish university.

4.1. Steps of the proposed system

Fig. 1 summarizes the seven-step PROMSET procedure and links each stage to its implementation and subsequent assessment in the empirical case study (Sections 4.2 and 5), providing a roadmap for the remainder of the paper.

Step 1. Definition of the evaluation problem and scope.

Evaluations should be conducted separately for each academic discipline. This separation by areas allows for controlling the bias associated with the area of knowledge, which significantly influences SET [14,30]. The first step is to define the specific knowledge area to be evaluated, which consists of a set of courses U. Each course $u \in U$ is associated with its corresponding academic field (for example, the marketing area includes courses such as Principles of Marketing, Marketing Management, and Market Research, among others). Then, the set of professors A is defined where each element $p_i \in A$ represents the evaluation of a professor who has taught a specific course from the set U. At this stage, the DM also defines ex ante the non-instructional bias to be evaluated and potentially mitigated, based on the institutional context and existing evidence in the literature.

Step 2. Definition of Criteria and Data Preprocessing.

A hierarchy of criteria is defined, structured into several dimensions. Each dimension contains a set of evaluation criteria G, where each criterion is measured through a question from the SET survey. Next, the data collected from students' responses is preprocessed. For each professor p_i and criterion g_j, the average rating Y^j is computed based on the individual scores given by the N students who completed the survey. This aggregation is performed using Eq. (9), where $S_{(l,p_i,g_j)}$ denotes the score given by student l to professor p_i on criterion g_j:

$$Y^j(p_i) = \frac{1}{N} \sum_{l=1}^{N} S_{(l,p_i,g_j)}. \tag{9}$$

Step 3. Identifying non-instructional bias using MLR.

To identify bias effects in SETs, this method proposes the use of MLR, a method frequently used in the literature about SET (see for example Zipser & Mincieli [31]). The goal is to estimate both the magnitude and the statistical significance of potential biases for each criterion g_j, using OLS with Robust Standard Errors (RSE).

The set of possible non-instructional biases that may affect criterion g_j is denoted by $X^j = \{x_1^j, x_2^j, \ldots, x_k^j, \ldots, x_z^j\}$. PROMSET evaluates the bias selected by the DM to determine its presence and magnitude across the evaluation criteria. From X^j, a set of regression coefficients $\beta^j = \{\beta_1^j, \beta_2^j, \ldots, \beta_k^j, \ldots, \beta_z^j\}$ is estimated, including the coefficient associated with the selected bias for each criterion.

The general form of the model for criterion g_j is shown in Eq. (10), where Y^j is the dependent variable. The term x_s^j represents the specific bias being addressed, with β_s^j as its corresponding coefficient. Other

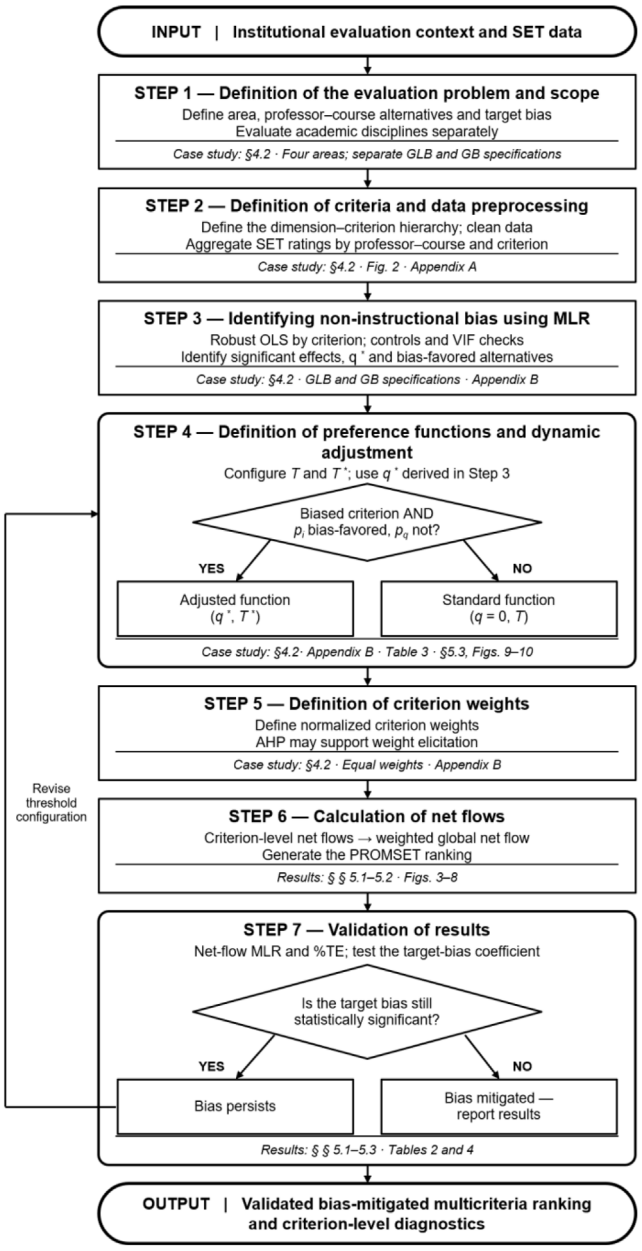


Fig. 1. PROMSET methodological flowchart and empirical implementation.

variables x_k^j are additional biases that are not directly mitigated but are included as statistical controls to improve the estimation of β_s^j. The term ε^j captures the random error:

$$Y^j = \beta_0^j + \beta_s^j x_s^j + \sum_{\substack{k=1 \\ k \neq s}}^{Z} (\beta_k^j x_k^j) + \varepsilon^j. \tag{10}$$

Potential multicollinearity among the explanatory variables should be assessed using variance inflation factors (VIF), particularly for the targeted bias variable. Then, once the regression coefficients are estimated, the statistical significance of β_s^j is evaluated using a predefined significance level α. The set of criteria affected by the bias x_s^j (i.e., where β_s^j is statistically significant) is defined as:

$$G_s = \left\{ g_j \in G \mid pvalue\left(\beta_s^j\right) \leq \alpha \right\}. \tag{11}$$

The proposed model mitigates one bias x_s^j at a time and applies only to those criteria where the associated coefficient β_s^j is significant. The

DM can choose to use either the estimated bias magnitude, $|\beta_s^j|$, or a conservative confidence-bound magnitude derived from its confidence interval, as a stricter mitigation, according to Eq. (12):

$$\beta^j_{s,lower} = \beta_s^j - t_{\frac{\alpha}{2},v} \cdot RSE\left(\beta_s^j\right), \quad \beta^j_{(s,upper)} = \beta_s^j + t_{\frac{\alpha}{2},v} \cdot RSE\left(\beta_s^j\right). \tag{12}$$

Next, for each criterion $g_j \in G_s$, a bias favorability vector $B_s^j = \left\{b(p_1)^j, b(p_2)^j, \ldots, b(p_i)^j, \ldots, b(p_n)^j\right\}$ is defined, where $b(p_i)^j$ is a binary variable indicating whether alternative p_i is being favored by the bias x_s^j. Its assignment depends on the type of variable x_s and the sign of β_s^j, in relation to the goal of criterion g_j (maximize or minimize). If x_s^j is a categorical variable (e.g., professor gender), then:

$$b(p_i)^j = \begin{cases} 1, & if\ p_i \in group\ favored\ by\ x_s^j \\ 0, & otherwise \end{cases}. \tag{13}$$

If x_s^j is a quantitative variable (e.g., students' expected grade), then:

$$b(p_i)^j = \begin{cases} 1, & If\ \left(\beta_s^j > 0 \wedge x_s^j(p_i) > R_s^j\right) \vee \left(\beta_s^j < 0 \wedge x_s^j(p_i) < R_s^j\right) \\ 0, & otherwise \end{cases}. \tag{14}$$

Here, R_s^j is a statistical threshold (mean, median, or another percentile) for the variable x_s^j within the full population.

Step 4. Definition of preference functions and dynamic adjustment.

Two preference functions are defined: a standard one, used when no mitigation is required, and an adjusted one, applied to criteria affected by significant bias.

The standard function F_j (Eq. (15)) sets the indifference threshold q_j equal to zero, corresponding to the V-shape function, meaning that any positive difference d_j must be reflected in the resulting preference Pr_j. The preference threshold T_j must be defined by the DM in accordance with institutional preferences.

$$F_j = \begin{cases} 0, & if\ d_j \le 0 \\ \dfrac{d_j}{T_j}, & if\ 0 < d_j \le T_j \\ 1, & if\ d_j > T_j \end{cases} \tag{15}$$

The adjusted function F_j^* introduces a bias-mitigated indifference threshold q_j^*, defined as $\left|\beta_s^j\right|$. The adjusted function F_j^* is applied when $g_j \in G_s$, and is defined as follows:

$$F_j^* = \begin{cases} 0, & if\ d_j \le |\beta_s^j| \\ \dfrac{d_j - |\beta_s^j|}{T_j^* - |\beta_s^j|}, & if\ |\beta_s^j| < d_j \le T_j^* \\ 1, & if\ d_j > T_j^* \end{cases} \tag{16}$$

The bias-mitigated preference threshold T_j^* defines the difference beyond which preference becomes complete in bias-activated comparisons and must satisfy $T_j^* > q_j^*$. Its selection can follow two complementary routes. First, it can be specified directly by the DM, as in the definition of the standard threshold T_j in Eq. (15). A natural institutional baseline is to set $T_j^* > T_j$, for example $T_j^* = 2T_j$, so that complete preference in bias-activated comparisons requires a larger observed difference than in standard comparisons. This stricter threshold prevents a bias-favored alternative from reaching full preference through small differences that may be attributable to the identified bias. The DM may adopt stricter or more lenient values, provided that $T_j^* > q_j^*$. Second, T_j^* can be empirically guided by Eq. (20) in Section 5.3, which provides, for each pair (T_j, q_j^*), a recommended value associated with the region of the parameter space where mitigation proved effective. Institutional preferences therefore condition both routes: directly in the first route, and through the standard threshold T_j and the selected value of q_j^* used as inputs in the second. In all cases, the selected configuration is subsequently validated in Step 7 through Eq. (18). If the bias remains statistically significant, the threshold configuration should be revised.

Once the preference functions are defined, pairwise comparisons are performed to compute $Pr_j(p_i, p_q)$, as shown in Eq. (17). The adjusted function is only used if the criterion is significantly biased and one of the alternatives p_i is favored by that bias, while the other alternative p_q is not:

$$Pr_j(p_i, p_q) = \begin{cases} F_j^*\left[d_j\left(p_i, p_q\right)\right], & if\ g_j \in G_s \wedge (b(p_i)^j = 1 \wedge b(p_q)^j = 0) \\ F_j\left[d_j\left(p_i, p_q\right)\right], & otherwise \end{cases} \tag{17}$$

Step 5. Definition of criterion weights.

The weight vector W is defined by assigning a specific weight to each evaluation criterion g_j. This is done by channeling the educational institution's preferences through the DM. The multicriteria AHP method can be used to support the definition of these weights (e.g., Do et al. [42], Thanassoulis et al. [49]). When several decision makers participate in the weighting process, their individual AHP judgments can be aggregated using the geometric mean to obtain a common set of criterion weights [55].

Step 6. Calculation of net flows.

Net flows are computed based on the previously defined preferences. For each alternative, both the net flow by criterion, $\varphi_j(p_i)$, and the overall net flow, $\varphi(p_i)$ are obtained, using Eqs. (4)–(8).

Step 7. Validation of results.

To verify whether the bias persists in the final evaluation, a new MLR model is estimated using $\varphi(p_i)$ defined in Eqs. (7)–(8) as the dependent variable for each alternative $p_i \in A$, where $A = \{p_1, \ldots, p_n\}$ denotes the full set of alternatives within the knowledge area:

$$\varphi(p_i) = \beta_0^{\varphi} + \beta_s^{\varphi} x_s^{\varphi} + \sum_{\substack{k=1. \\ k \ne s}}^{Z} \beta_k^{\varphi} x_k^{\varphi} + \varepsilon^{\varphi}. \tag{18}$$

The result will indicate whether β_s^{φ} is statistically significant. If it is not, the targeted bias is considered to be no longer statistically detectable at the predefined significance level for the purposes of the PROMSET validation procedure. Finally, based on $\varphi\left(p_i\right)$ and $\varphi_j(p_i)$ a global ranking and a detailed per-criterion analysis are generated to support decision-making within the institution.

As an additional ranking diagnostic, the discriminating power of the ranking can be assessed by calculating the percentage of technical ties (%TE), which compares the relative differences in net flow between alternatives ranked in consecutive positions r and $r+1$. Two alternatives are considered technically tied if their relative difference in net flow is less than a predefined threshold t. The total number of ties is computed by summing all comparisons between consecutively ranked alternatives, as defined in Eq. (19). A technical tie means that the model does not provide enough evidence to prefer one alternative over the next, so their relative order is not informative for decision-making. Conversely, the ranking distinguishes between two consecutive alternatives when their relative net-flow difference is greater than or equal to t, because their order then reflects a meaningful difference in aggregated performance. If the denominator in the relative difference is zero, the comparison is made using the absolute difference between both net flows.

$$\%TE = \frac{\sum_{r=1}^{n-1} 1\left(\dfrac{\left|\varphi\left(p_i^r\right) - \varphi\left(p_q^{r+1}\right)\right|}{\max\left(\left|\varphi\left(p_i^r\right)\right|, \left|\varphi\left(p_q^{r+1}\right)\right|\right)} < t\right)}{n-1} \cdot 100. \tag{19}$$

4.2. Validation of the proposed system: faculty evaluation using real SET data

Step 1. Definition of the evaluation problem and scope.

The validation was carried out using real data from surveys conducted at a Spanish university. Four knowledge areas were evaluated: quantitative methods, finance, economics, and international relations.

A total of 879 courses were included across the four areas analyzed. Each course is defined as a unique combination of a professor teaching a specific subject to a particular group of students. Professors could be assigned to one or more courses. To structure the analysis, four

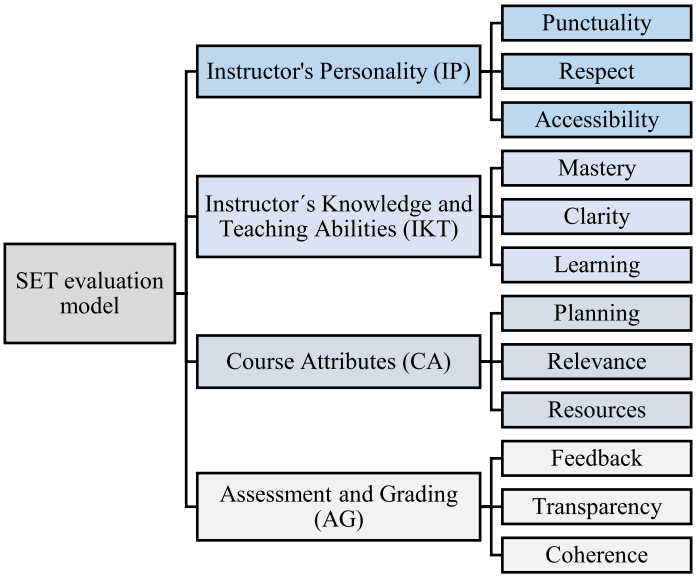


Fig. 2. Hierarchical structure of evaluation criteria in the SET model.

sets A were defined, each containing a specific number of alternatives corresponding to a course within each knowledge area. In quantitative methods, 223 courses were evaluated; in finance, 320; in economics, 190; and in international relations, 146. The evaluation scale used in the SETs ranges from 1 to 10, with 10 being the highest score assigned by a student.

The objective is to generate preference-ordered rankings for each academic area, based on the professors teaching the courses. A single professor may appear more than once if they taught multiple courses. In all cases, the first position in the ranking corresponds to the most preferred alternative, with subsequent positions reflecting a descending order of preference.

The target bias should be defined ex ante by the DM based on empirical evidence from the institution and existing literature, which has documented various non-instructional factors that can distort SET ratings. Following this criterion, the empirical validation includes the mitigation of GLB, a well-documented phenomenon in the literature [20,56]: professors who give higher grades tend to receive better teaching evaluations. Additionally, to validate the system's effectiveness against biases associated with categorical variables, GB, linked to the professor's gender [57], was evaluated in each knowledge area and mitigated wherever statistically significant, following the same procedure.

Finally, the results are compared with those produced by the traditional method used by the university, based on a ranking generated from the average score of a single SET question ("Overall evaluation of the professor"). Additionally, the PROMSET outcomes are contrasted with rankings obtained using classical PROMETHEE II under standard threshold parametrization, in order to isolate the structural effect of bias activation.

Step 2. Definition of criteria and data preprocessing.

Fig. 2 presents the evaluation criteria hierarchy, which includes four dimensions divided into twelve criteria. These dimensions are defined based on factors identified in the literature on faculty evaluation through SETs [11,58].

The Instructor's Personality (IP) dimension evaluates the professor's interaction with students and classroom attitude; Instructor's Knowledge and Teaching Abilities (IKT) measures subject–matter expertise and pedagogical skills; Course Attributes (CA) focuses on planning and the resources used in teaching; and the Assessment and Grading (AG) dimension addresses feedback, transparency, and consistency in the evaluation process. A detailed list of the SET questions associated with each criterion is provided in Appendix A.

During the data cleaning process, records with incomplete or inconsistent values were removed, and groups with fewer than ten survey responses were excluded to ensure the reliability of the evaluation. After cleaning, the final dataset included 45,357 surveys. Then, using Eq. (9), the average rating was calculated for each of the 879 courses taught by professors, according to each evaluation criterion.

Step 3. Identifying non-instructional bias using MLR.

As stated in Step 1, GLB and GB were selected as the two bias variables to be mitigated in the empirical validation, each playing the role of x_s^j. In this application, GLB was proxied by the average course grade recorded for each professor–course evaluation, since it provides an observable course-level measure of the grade-related component potentially reflected in SET responses. GB was operationalized using the gender category recorded in the university's administrative records. This choice was based on three considerations. First, both biases are among the most extensively studied non-instructional influences documented in the SET literature. Second, both biases were found to be statistically significant in several knowledge areas of this institutional context, making them suitable cases for validating the proposed mitigation mechanism. Third, they represent the two types of bias variables contemplated by PROMSET: GLB is continuous, whereas GB is categorical.

The remaining variables in x^j, including academic category, professor age, nationality, research level, full-time dedication, year of appointment, and first year at the institution, were included as statistical controls. When each bias was estimated, the other was also retained as a control. Thus, the estimated GLB effect was conditional on professor gender, whereas the estimated GB effect was conditional on average course grade. The two biases were nevertheless mitigated in separate PROMSET specifications, each with its own bias-specific parameters. Their inclusion allows the estimated effect of the mitigated bias to be isolated conditional on other potential non-instructional influences identified in previous research, even if these factors were not consistently significant across all knowledge areas in the empirical context analyzed.

Following Eq. (10), the coefficient of the bias variable selected for mitigation, x_s^j, was estimated separately for each of the twelve evaluation criteria g_j and each knowledge area. All VIF values for the bias variables and controls were below 5 across the four knowledge areas. Statistical significance was assessed at $\alpha = 0.05$, following Eq. (11). For significant coefficients, the value used for mitigation corresponds to the upper confidence bound of the absolute estimated bias effect, following Eq. (12), so that the adjusted preference function discounts

Table 2
Significance of GLB across knowledge areas.

	Quantitative methods	Economics	Finance	International relations
Traditional Method				
GLB p-value	0.008	<0.001	0.020	<0.001
%TE	91.9	88.4	96.2	91.0
PROMSET				
GLB p-value	0.435	0.112	0.203	0.060
%TE	37.4	29.1	46.7	24.8
Ranking Comparison				
Kendall Tau (p-value)	0.80 (<0.001)	0.80 (<0.001)	0.84 (<0.001)	0.79 (<0.001)
Average absolute variation	15.9	13.5	17.9	11.4

Note. %TE for differences smaller than 1%, according to Eq. (19).

a conservative estimate of the bias magnitude. The bias-favorability indicators were then defined according to Eqs. (13) and (14).

Step 4. Definition of preference functions and dynamic adjustment.

Regarding the parametrization of the preference functions in the empirical validation, the preference-driven route described in Section 4.1 was adopted. The standard preference threshold was set to $T_j = 1.5$ for all criteria. This choice is grounded on three complementary considerations. First, $T_j = 1.5$ represents 16.7% of the evaluation scale range (1–10), which falls within the 10%–20% interval recommended in applied PROMETHEE practice when the decision-maker does not specify a more precise value [12,54]. Second, from an institutional perspective, this threshold implies that complete preference between two professors on a given criterion is only reached when the difference in their average student ratings is at least 1.5 (for instance, the difference between a 7.0 and an 8.5 on a 1–10 scale), a requirement that is demanding yet practically meaningful for academic decision-making. Third, this initial choice is validated a posteriori by the sensitivity analysis reported in Section 5.3, which shows that among the parameter configurations where GLB was effectively mitigated, T shifts toward lower values (Wilcoxon, $p < 0.001$; Fig. 9), confirming that $T_j = 1.5$ lies within the region of the parameter space associated with successful bias mitigation. Using a uniform threshold across all items further avoids criterion-specific threshold effects unrelated to bias mitigation.

For criteria affected by the bias, the preference-driven option adopted in the main validation was to set the stricter bias-mitigated preference threshold to $T_j^* = 2T_j$. This required complete preference in bias-activated comparisons to be supported by a larger observed difference than in standard comparisons. The preference functions were then specified according to Eqs. (15) through (17).

Across the four knowledge areas, the GLB parameter configurations differed in both the number of affected criteria and the magnitude of q^*. GLB affected 5 of the 12 criteria in finance, compared with 7 in quantitative methods, 10 in international relations, and 11 in economics. At the criterion level, learning and relevance were identified as affected by GLB in all four areas, whereas punctuality was unaffected in all cases and mastery was affected only in economics. Among affected criteria, q^* ranged from 0.4–0.5 in finance, 0.4–0.7 in quantitative methods, 0.6–1.4 in international relations, and 0.5–1.1 in economics. Detailed criterion-level parameter values are provided in Tables B.1 to B.4 in Appendix B.

Step 5. Definition of criterion weights.

To conduct a more objective validation of the model's mitigation effect, equal weights were assigned to all criteria (8.33% each). This avoids any influence of weighting on the statistical significance of the bias in the final ranking.

Steps 6 and 7. Net flows and validation.

The calculation of net flows and the validation of the resulting rankings are presented in Section 5. Step 7 uses Eq. (18) to validate bias mitigation, with %TE as an additional ranking diagnostic.

5. Results

This section reports the empirical results, following the structure of the methodology. Section 5.1 presents the general results for the four knowledge areas, corresponding to Steps 6 and 7 of PROMSET. Section 5.2 provides a more detailed analysis for the quantitative methods area, while Section 5.3 assesses the sensitivity of the Step 4 preference-function parameters using the Step 7 validation procedure.

5.1. General results: Steps 6 and 7

Following Step 6, the criterion-level and global net flows were computed to generate the PROMSET rankings. Following Step 7, the resulting global net flows were used to assess whether the target bias remained statistically significant and to evaluate the discriminating power of the rankings through %TE.

For the calculation of technical ties, the tolerance threshold was set to $t = 0.01$, so two consecutive alternatives were considered tied when their relative difference was less than 1%. This value was fixed a priori as a practical tolerance. The threshold is used only to compute the %TE diagnostic metric and does not affect the preference functions, net flows, or final rankings. Since the same tolerance is applied uniformly across all methods and knowledge areas, %TE comparisons are made on a common basis.

5.1.1. Mitigation of grade leniency bias

To validate the effectiveness of the model in mitigating GLB, Table 2 presents a comparison of the rankings obtained through PROMSET and the traditional approach across the four knowledge areas.

First, the p-values from the regression analysis show a considerable decrease in the statistical significance of GLB when using PROMSET. While the traditional ranking yields statistically significant results ($p < 0.05$) in all areas, PROMSET produces p-values above the significance threshold, suggesting that the bias has been mitigated.

Second, the proposed method significantly reduces the %TE in all areas. Whereas the traditional approach shows values above 88%, PROMSET lowers this to a range between 24.8% and 46.7%, indicating greater discriminative power of the model.

In addition, the Kendall Tau correlation coefficient [59] reveals a statistically significant positive association between both rankings, with concordance values ranging from 0.79 to 0.84. Furthermore, the average absolute difference in positions between rankings varies from 11.37 to 17.9 positions. This suggests that, while a general ranking structure is preserved, the proposed model introduces specific changes in professor positions as a result of the adjustment for non-instructional biases.

These findings reinforce the idea that the model not only maintains a degree of consistency with the traditional order but also introduces meaningful adjustments related to bias mitigation. Given these positive outcomes, a more detailed analysis is conducted for the quantitative methods area.

Table 3
GLB p-values under alternative PROMETHEE II specifications.

Threshold calibration	Quantitative methods	Economics	Finance	International relations
Standard thresholds	0.027	<0.001	0.009	<0.001
Thresholds incorporating estimated bias	0.037	<0.001	0.027	<0.001

Note. In all preference functions, "Standard thresholds" set $q = 0$ and use the standard preference threshold T; "Thresholds incorporating estimated bias" set q^* and T^* without comparison-dependent activation.

To assess sensitivity to the tolerance threshold, %TE was recalculated at 0.5% and 5%. For the 0.5% threshold, %TE values (Traditional/PROMSET) were as follows: quantitative methods 78.8/24.3, economics 76.7/20.1, finance 89.0/27.9, and international relations 71.7/11.0. At the 5% threshold, the corresponding values increased to 99.1/79.7, 97.9/75.7, 99.4/87.5, and 97.9/65.5, respectively. In all cases, PROMSET showed lower %TE values than the traditional method.

Table 3 reports the statistical significance (p-values) of GLB in the net flows obtained under two PROMETHEE II benchmark configurations: standard thresholds and thresholds incorporating the estimated bias magnitude into their parametrization. In both cases, GLB remains statistically significant across all four knowledge areas ($p < 0.05$). These findings indicate that neither the standard threshold specification nor the bias-informed parametrization eliminates the structural effect of GLB on the resulting net flows and ranking.

5.1.2. Mitigation of gender bias

Following the same Step 6 and Step 7 procedure, GB mitigation was evaluated independently within each knowledge area. In the traditional specification, GB was statistically significant in quantitative methods, finance, and international relations, so mitigation was applied in these three areas. The results are reported in Table 4. When PROMSET is applied following the same procedure separately by knowledge area, these effects are no longer statistically detectable at the 5% level.

Likewise, the percentage of technical ties decreases substantially under PROMSET across the three areas. This finding complements the previously reported evidence on GLB and further supports the model's effectiveness in addressing different types of non-instructional biases, including those associated with categorical characteristics such as gender, thereby reinforcing its validity and practical applicability.

As an additional check for potential cross-bias effects, the non-targeted bias was reassessed after each mitigation procedure. In both directions, its significance pattern remained unchanged: after mitigating GLB, GB remained significant in the same three areas and non-significant in economics, while after mitigating GB, GLB remained significant in all three areas where GB mitigation was applied.

5.2. Detailed results: quantitative methods

5.2.1. Mitigation of grade leniency bias

Fig. 3 shows a comparison between the rankings obtained using the traditional method and PROMSET. The diagonal line represents equality between methods. Professors located below this line improved their ranking after applying PROMSET, while those above experienced a decline. A general positive correlation can be observed, although there are deviations from the equality line. This confirms that position changes do occur, particularly in the middle segments of the ranking, where the most significant reordering is observed.

Fig. 4 illustrates the magnitude of individual position changes resulting from the application of PROMSET, in comparison to the traditional ranking. Greater dispersion is observed in the middle quartiles (Q), where variations occur more frequently. The Root Mean Square Error (RMSE) reaches 26.5 in Q2 and 21.1 in Q3, exceeding the values

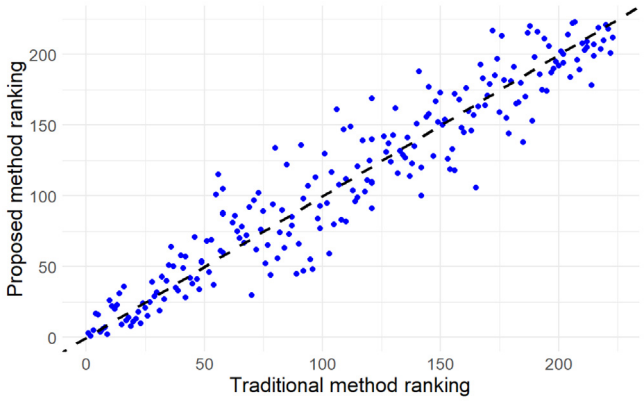


Fig. 3. Comparison of rankings between the proposed and traditional methods.

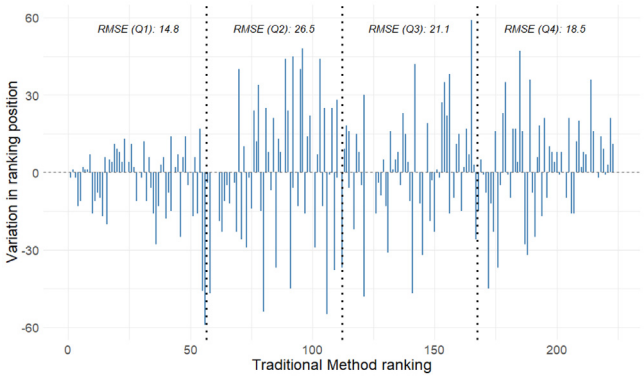


Fig. 4. Magnitude of changes in ranking position.

recorded at the extremes: 14.8 in Q1, corresponding to the top positions in the ranking, and 18.5 in Q4, associated with the lowest positions.

These results suggest that the traditional ranking may be more sensitive to undesirable factors, such as biases, particularly in the middle ranges. Consequently, PROMSET introduces more substantial adjustments in these segments. This is a particularly interesting result, as previous studies have pointed out that bias in SET calls into question the validity of their outcomes, especially in mid-range evaluations (see, for example, Arroyo-Barrigüete et al. [22]). Therefore, this finding is consistent with what would be expected based on earlier research.

This variability in rankings is further analyzed in Fig. 5, which visualizes the movements between quartiles. Each cell indicates the number of faculty evaluations that shift from one quartile to another, and the average grade associated with GLB for each group is shown in parentheses. This visualization reinforces the previous observations by highlighting that the changes are mainly concentrated in the middle quartiles.

Additionally, evaluations that move down a quartile under the proposed method tend to have, on average, higher GLB scores than those that move up. This suggests that their original positions may have been favored by a bias associated with average grade in the traditional ranking. For example, seven evaluations originally placed in Q1, with an average GLB score of 7.14, drop to Q2, while eight evaluations from Q2 move up to Q1, with an average score of 6.22.

Beyond bias mitigation, PROMSET also provides valuable support for decision-making, as it allows for disaggregated interpretation of results based on the established criterion hierarchy. Fig. 6 shows how the overall net flow is decomposed for each course evaluation into the model's dimensions: IKT, IP, CA, and AG.

Table 4
Significance of GB across knowledge areas.

	Quantitative methods		Finance		International relations	
	P-value	%TE	P-value	%TE	P-value	%TE
Traditional method	0.007	91.9	<0.001	96.2	<0.001	91.0
PROMSET	0.768	40.1	0.685	44.5	0.458	43.6

Note. The significant GB effect indicates a bias in favor of female professors in quantitative methods and in favor of male professors in finance and international relations. Economics is not reported because GB was not statistically significant in the traditional specification and therefore no mitigation was required.

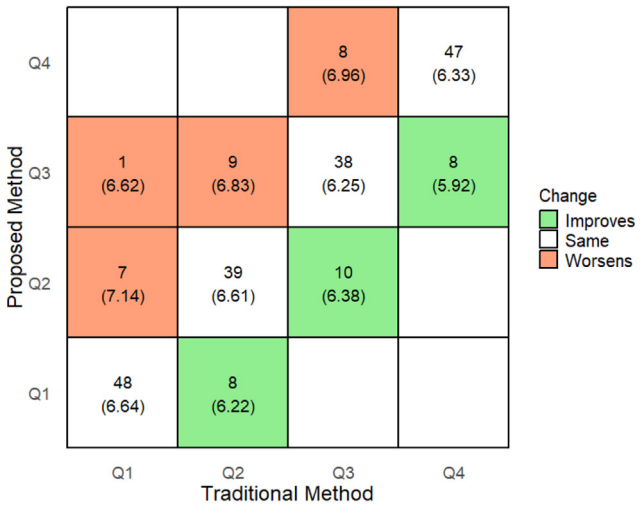


Fig. 5. Mobility across ranking quartiles.

In this context, positive values indicate a stronger preference, values near zero reflect indifference and negative values represent a lower preference. This decomposition helps identify differentiated profiles across evaluations. For instance, some observations show a positive net flow driven primarily by a single dimension, while others display a more balanced contribution across dimensions. There are also cases with contrasting performance, where an evaluation is preferred in some areas but disadvantaged in others.

The analysis of combined positioning across strategic dimensions offers an additional perspective on teaching performance. Fig. 7 focuses specifically on the IKT and IP dimensions and shows the relative positioning of each evaluation based on these two key variables. In the first quadrant (I), evaluations correspond to professors whose courses received a positive net flow in both dimensions, indicating preference in both IKT and IP. The second quadrant (II) includes cases with a positive net flow in IP and a negative flow in IKT. The third quadrant (III) groups professors with low preference in both dimensions, while the fourth quadrant (IV) reflects cases with high preference in IKT and low preference in IP.

This representation provides a complementary reading to the overall analysis, as it allows the identification of different teaching profiles based on combined performance. It also reinforces the model's usefulness as a diagnostic tool to support decision-making. For example, from an overall perspective, a professor from quadrant II and another from quadrant IV might exhibit a very similar net flow, and therefore a comparable position in the ranking. However, depending on which of the two dimensions is more relevant from an institutional standpoint, one may be preferred over the other. This type of analysis is useful for tailoring professional development to the specific needs of each instructor. In other words, beyond simply providing a ranking, PROMSET offers a high level of granularity to support informed decision-making.

5.2.2. Mitigation of gender bias

Following the same approach applied to GLB, Fig. 8 examines the changes produced by PROMSET when mitigating GB in this area. Fig. 8(a) shows, for each evaluation, the change in ranking position between the traditional ranking and the ranking obtained with PROMSET, together with the RMSE for each quartile; the largest ranking shifts occur in the middle quartiles. Fig. 8(b) shows mobility across quartiles, with the proportion of female professors reported in each cell. Fig. 8(c) shows the distribution of dimension-level net flows (IKT versus IP) by professor gender.

The pattern observed in Fig. 8(b) is consistent with the bias that PROMSET is intended to mitigate: movements do not exceed one quartile, and most of the 25 evaluations that move down a quartile correspond to female professors, whereas only a minority of those that move up do. Since the bias detected in this area favored female professors, this is the expected effect of the mitigation. As in the GLB analysis, the dimension-level representation in Fig. 8(c) reveals different performance profiles from the joint positioning on IKT and IP. Male and female professors are similarly distributed across the quadrants, with no significant association between gender and quadrant membership ($\chi^2(3) = 0.363$, $p = 0.948$). The remaining pairwise dimension combinations showed the same pattern, suggesting no clear systematic gender-based differentiation after mitigation.

5.3. Sensitivity analysis of Step 4 parameters and Step 7 validation

A sensitivity analysis was conducted to assess the persistence of GLB in the resulting net flows. A total of 11,016 evaluations were performed using different parameter configurations, with the aim of analyzing the statistical significance of the bias through the resulting p-values.

In each knowledge area, 2754 possible combinations were analyzed. These were generated by varying three parameters: the preference threshold T, the bias-mitigated preference threshold T^*, and the bias-mitigated indifference threshold q^*. In the main validation, q^* was defined from the estimated bias magnitude. For sensitivity purposes, however, q^* was treated as a free calibration parameter, so that varying it emulates alternative magnitudes of the bias-mitigated indifference threshold. The values of these parameters were varied within the range [0, 4], using increments of 0.2, and ensuring that in all cases $q^* \neq 0$, $T^* - q^* \geq 0.5$ and $T - q \geq 0.5$. Since $q = 0$ in the standard preference function, the latter condition implies $T \geq 0.5$. The range [0, 4] was chosen to include values below and above the baseline threshold $T_j = 1.5$ used in the main validation, thereby assessing the robustness of the mitigation results to this initial choice. The 0.2 increment provided a sufficiently fine yet computationally manageable grid, while the constraint $q^* \neq 0$ ensured that the adjusted function retained its bias-discounting mechanism. $T^* - q^* \geq 0.5$ prevented near-step preference functions, and $T - q \geq 0.5$ imposed the same minimum partial-preference width on the standard function, preventing the same near-step behavior in comparisons where bias mitigation is not activated.

Each configuration was evaluated using Eq. (18) to verify whether the bias remained statistically significant under each scenario. This procedure made it possible to identify the parameter ranges in which the mitigation is most effective, reinforcing the model's robustness across different adjustments.

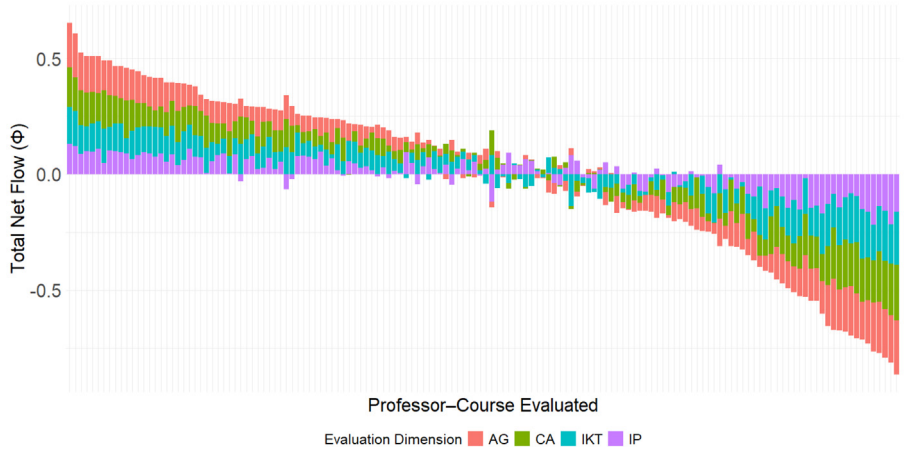


Fig. 6. Dimension-level contributions to total Net Flow (Φ).

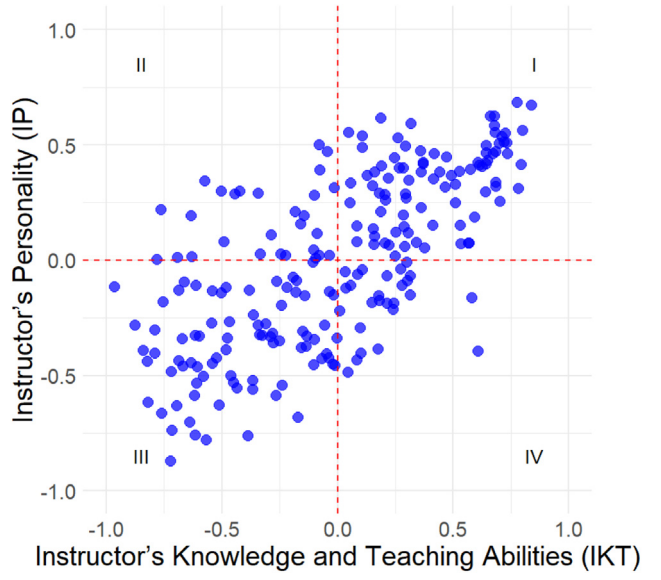


Fig. 7. Comparison of net flows: IP vs IKT. Four quadrants representing performance profiles.

Fig. 9 shows the distributions of the parameter values q^*, T^*, and T, distinguishing between models where the GLB was no longer significant after mitigation (blue group) and those where the bias remained significant (red group). In other words, the figure presents the values used for each parameter, separated according to the statistical significance of GLB in the resulting model. The results show that when the bias has been mitigated, the values of T^* and q^* tend to be higher, while T shifts toward lower values. These differences, observed within each parameter when comparing bias-mitigated (blue) and unmitigated (red) models, are statistically significant (Wilcoxon, $p < 0.001$). This indicates that bias-free models tend to adopt different parameter configurations compared to those where the bias persists.

Within the subset of models where the bias is not significant (blue group), a separate analysis was conducted to compare the values taken by the three parameters. This triple comparison, limited to the cases in which the bias was mitigated, revealed significant internal differences. The value of T^* is significantly higher than T, and T, in turn, is greater than q^* (Wilcoxon, $p < 0.001$; adjusted using Holm's method). In addition, when directly comparing the difference $T^* - q^*$ to the value of T, it was found that this difference represents, on average, 63.6% of T (Wilcoxon, $p < 0.001$).

This result indicates that although T^* is greater than T, the increase in q^* reduces the effective range for partial preference. On average, the difference $T^* - q^*$ accounts for only 63.6% of the value of T, implying a more moderate preference level in cases affected by bias. Thus, the model not only mitigates the bias but also limits its influence on the final ranking.

This dynamic is more clearly visualized in Fig. 10, which presents a three-dimensional scatter plot showing the evaluated combinations of T, T^*, and q^*. Combinations in which the evaluated bias is not significant in any of the areas are shown in blue, while those where the bias persists in at least one area are shown in red. To facilitate interpretation, direct annotations identify the bias-mitigated and bias-persistent regions, while the gray plane represents the empirically fitted Eq. (20):

$$T^* = 3.0331 + 0.8916T - 1.2840q^*, \quad \forall T > 0, \quad 0 \le q^* < T^*. \tag{20}$$

Of the 2754 parameter combinations evaluated, 328 (11.9%) mitigated GLB simultaneously in all four knowledge areas. These non-significant combinations are concentrated around the fitted plane, suggesting that the relationship defined by the plane can serve as an empirical reference for threshold selection. In particular, it enables the DM to calculate a candidate value of T^* based on the previously defined value of T and the estimated value of q^*.

To test the practical applicability of the fitted plane equation, the values of T^* were calculated for each knowledge area using Eq. (20). The results demonstrate that, when these adjusted values are applied, GLB is no longer statistically significant in the net flows. Specifically, the p-values are 0.670 in quantitative methods, 0.343 in finance, 0.153 in economics, and 0.065 in international relations. These findings support the usefulness of the fitted relationship within the present dataset. Since the equation and this subsequent check are based on the same institutional dataset, the results should be interpreted as evidence of internal consistency rather than external validation.

An analogous sensitivity analysis was conducted for GB in the three knowledge areas where it was significant, using the same parameter grid. GB was simultaneously mitigated across all three areas in 828 configurations (30.1%). Successful configurations showed the same general directional pattern observed for GLB in Fig. 9, with higher values of q^* and T^* and lower values of T. These results indicate that GB mitigation is compatible with a broader region of the parameter space. However, this comparison should be interpreted with caution, since GB was evaluated across three areas and GLB across four.

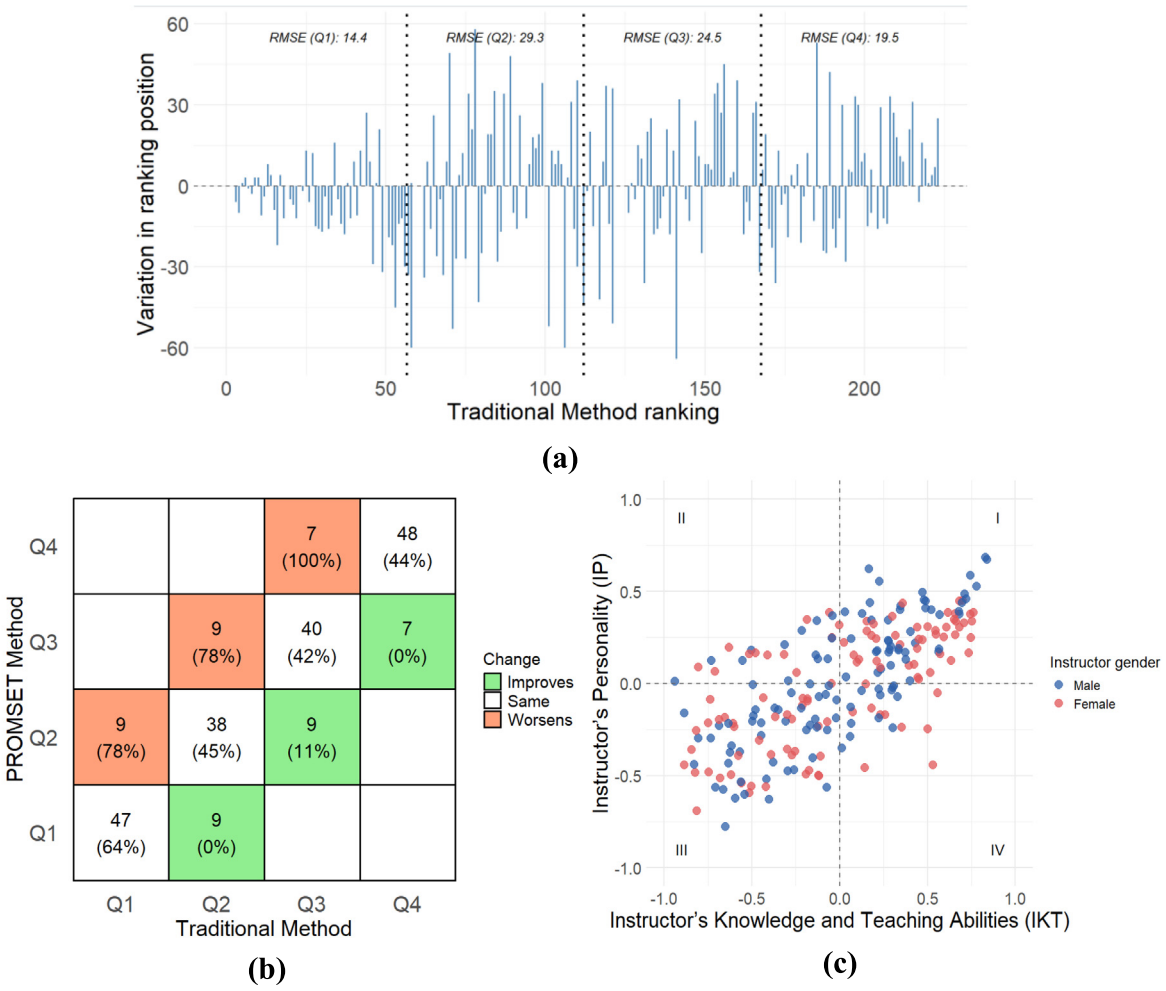


Fig. 8. Effects of GB mitigation in quantitative methods: (a) Magnitude of changes in ranking position, (b) Mobility across ranking quartiles and female representation, and (c) IP–IKT net-flow profiles by instructor gender. Percentages in parentheses in panel (b) indicate the proportion of female professors.

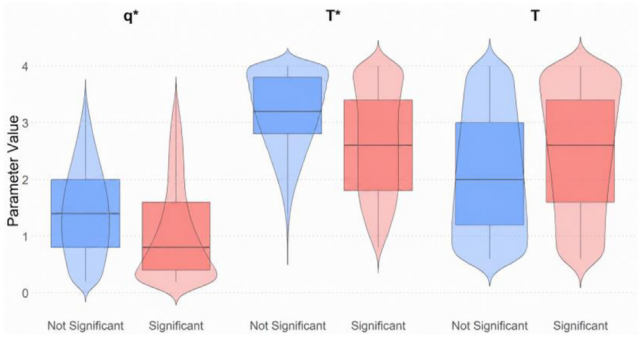


Fig. 9. Distribution of parameters according to the significance of the GLB.

As an additional specification–sensitivity check, the criterion-level bias models were re-estimated without the control variables for each knowledge area and bias. The number of statistically significant effects changed from 33 to 37 for GLB and from 35 to 37 for GB, with 31 effects remaining significant under both specifications in each case; no coefficient significant in either specification changed sign. The resulting changes in q_j^* varied across areas. When PROMSET was recalculated for the seven scenarios using the reduced specification, the bias was no longer statistically significant in any case ($\alpha = 0.05$). The controls are nonetheless retained in the main specification to estimate the targeted bias conditional on other potential non-instructional influences.

6. Discussion

This study proposes PROMSET as an evaluation system incorporating a structural extension of the PROMETHEE II methodology aimed at identifying and mitigating biases in survey-based evaluation systems. Specifically, the model has been validated in the context of non-instructional biases observed in SET surveys. By integrating econometric bias detection with comparison-dependent calibration of preference thresholds, PROMSET introduces a bias-sensitive activation mechanism within the outranking procedure.

Empirical validation was conducted using SET data from 879 courses across four knowledge areas. The assessment employed 12 criteria grouped into four dimensions and examined two prominent non-instructional biases, GLB and GB, modeled through continuous and categorical variables, respectively. This design enabled testing PROMSET's mitigation capability under diverse empirical conditions, including multiple academic domains and both continuous and categorical bias structures. Specifically, across all four knowledge areas, GLB coefficients, originally significant, became statistically nonsignificant ($p > 0.05$) post-implementation of PROMSET. Similarly,

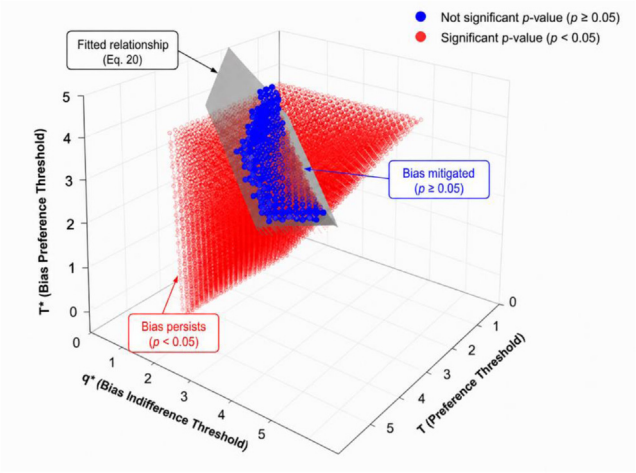


Fig. 10. Three-dimensional distribution of T, T^*, and q^* according to GLB significance, with annotated mitigation regions and the empirical relationship from Eq. (20).

GB no longer reached statistical significance at the 5% level in the areas where it had been detected in the traditional approach. These results provide evidence of PROMSET's capability to mitigate biases identified through both continuous and categorical variables. Consequently, integrating MLR models systematically with the PROMETHEE II extension yields a structured adjustment mechanism that reduces undeserved advantages associated with inflated grades or gender-related distortions.

The sensitivity analysis suggests that GB can be mitigated across a broader range of parameter configurations than GLB. Consistently, under the main parametrization, post-mitigation p-values for GB are also farther from the significance threshold than the corresponding GLB p-values. One possible explanation lies in how bias-favored alternatives are identified: for GB, this status is determined directly by group membership, whereas for GLB it is defined using a cutoff. This difference may make the calibration required to mitigate GLB more restrictive. However, this comparison should be interpreted with caution, since mitigation was evaluated across three areas for GB and four for GLB.

Importantly, bias mitigation in PROMSET relies on a comparison-dependent activation mechanism rather than on threshold parametrization alone. As shown in Table 3, GLB remains statistically significant under PROMETHEE II even when thresholds incorporate the identified bias magnitude. The difference between the two approaches is therefore structural rather than merely parametric.

In addition, PROMSET enhances ranking granularity by leveraging the well-established discriminative power of PROMETHEE II net flows reported in prior studies [60]. The proportion of ties decreases substantially, falling between 24% and 47%, compared to over 88% using the classical approach. Moreover, bias mitigation leads to the greatest ranking rearrangement in middle positions, highlighting that GLB particularly compromises evaluation reliability in this segment. This directly addresses a key concern in the use of SET: their lack of reliability, especially in mid-range rankings, as noted by Arroyo-Barrigüete et al. [22].

Furthermore, the multicriteria assessment across four distinct dimensions uncovers differentiated instructor profiles. For instance, instructors excelling in IKT, but demonstrating lower performance in IP, can be identified, thus suggesting targeted educational interventions. This nuanced analysis confirms PROMSET's diagnostic value, extending beyond mere global rankings.

Various studies have previously addressed biases in SET, yet they typically rely on partial approaches. One line of research focuses on bias detection but lacks integrated mitigation strategies. Another promotes ex-ante mitigation by introducing bias-awareness prompts before data collection [27,28]. However, Kim et al. [29] show that such interventions may also produce unintended effects, including lower ratings for some groups. A middle-ground approach, as seen in Arroyo-Barrigüete et al. [30] and Zipser and Mincieli [31], applies OLS models to isolate bias effects and uses residuals as adjusted performance indicators. However, while these methods improve the equity of SET, they do not appear to achieve full bias removal. In contrast, PROMSET offers a comprehensive and systematic evaluation model that enables the integrated, post-hoc identification and mitigation of biases while organizing all SET data into institutionally aligned, comparable rankings. These findings are consistent with prior evidence on non-instructional biases in SET. As Berezvai et al. [20] documented, ratings can increase by up to 0.4 points for each additional unit of average grade. Similarly, after applying bias mitigation, GLB often loses statistical significance, and previously favored instructors tend to drop in rank (see Fig. 5).

Previous MCDM research on faculty evaluation has relied predominantly on AHP and its fuzzy variants for criterion weighting [1,42]. AHP has also been used to examine how different stakeholders value evaluation criteria; Hadad et al. [13], for example, compared the importance assigned to these criteria by students and faculty. While these techniques capture decision-maker preferences with considerable rigor, they require pairwise judgments to be elicited and the resulting comparison matrices to be checked for consistency. In the present study, equal weights were assigned to the twelve criteria to eliminate any confounding effect that differential weighting might exert on bias significance. This design choice supports the interpretation that the observed bias reduction arises from PROMSET's internal adjustment mechanism rather than from an artifact of the weighting scheme. Nevertheless, the framework remains fully compatible with AHP or similar hierarchical procedures, so that institutions can incorporate strategic priorities through differentiated weights.

Outranking approaches such as PROMETHEE have recently gained traction [48], yet they have not been applied exclusively to raw SET data to date. In the broader MCDM literature, Thanassoulis et al. [49] applied AHP and DEA to SET data, but their model focuses on multicriteria aggregation and does not explicitly address the detection or mitigation of well-documented non-instructional biases [4]. By contrast, PROMSET integrates econometric bias identification through MLR with an outranking aggregation that uses dynamically calibrated preference functions. This combination leverages the full descriptive power of SET surveys while ensuring faculty rankings that are demonstrably free from systematic distortions.

It is critical to clarify that this work does not advocate for SET as the primary instrument for instructor evaluation, a widely debated subject in educational research [21]. However, given SET's prevalent institutional usage [4], this study provides universities with a valuable tool to more appropriately adjust evaluation results, thereby supporting decision-making processes that are better aligned with institutional objectives.

When PROMSET is used to support high-stakes personnel decisions, its application should be embedded within institutional mechanisms for transparency and review. Because the procedure is reproducible and its parameters are explicit, the resulting ranking position can be reported together with the dimension-level net flows that help explain it. Documentation can also include the input data, the classification of alternatives as bias-favored, the parameter values used, and the comparisons in which mitigation was activated. This traceability would facilitate review or appeal procedures that can verify both the data and the application of the method. Bias estimates and calibration should be reviewed periodically, as they may vary across contexts and over time. Regarding potential strategic adaptation, PROMSET limits the advantage that could arise from higher grades by discounting this effect in the affected comparisons. Calibration at the knowledge-area level reduces an individual instructor's ability to influence the resulting thresholds. Nevertheless, PROMSET should be used as one component of a broader

evaluation process that also considers other relevant aspects of professor performance, rather than as an automatic rule for individual personnel decisions.

Although the empirical validation focuses on SET data, the proposed mechanism is not inherently restricted to the educational context. The identification of bias and its structural incorporation into the aggregation procedure make PROMSET methodologically transferable to other survey-based evaluation systems in which systematic biases affect ranking outcomes.

While PROMSET is presented from the perspective of a single decision maker, its structure does not require the parameters to be defined unilaterally. PROMSET can also incorporate preferences agreed upon by various stakeholders through a common set of weights and thresholds. When different groups maintain their own evaluations or preferences, one possible adaptation would be to analyze each perspective separately and then integrate the results using PROMETHEE-GDSS [61]. Sensitivity analysis could provide an empirical basis for deliberation on the thresholds, while Step 7 would allow for the validation of each specification with respect to the target bias. In this context, divergent stakeholder preferences should also be distinguished from strategically biased reporting. The identification strategy captures systematic components associated with observable covariates and therefore cannot determine whether the observed patterns reflect unintentional response tendencies or deliberate reporting behavior. PROMSET acts on the observed effect rather than on the underlying behavior that generates it.

From a statistical standpoint, the significance tests in Step 3 are conducted separately by knowledge area and bias specification. These tests serve as a filter to determine which preference functions are adjusted, rather than as final inferential results. Given this screening role, no multiple-comparisons correction was applied to the 12 criterion-level tests within each area and specification. Tests across knowledge areas or bias types were not combined into a single family because knowledge areas are analyzed using distinct course subsets, while GLB and GB are estimated under separate specifications. This choice involves an explicit trade-off between sensitivity and protection against false positives: a strict correction could reduce sensitivity to moderate associations that, spread across several criteria, may jointly contribute to bias detectable in the aggregate net flow, whereas not applying a correction may result in adjusting a criterion that is not actually affected. To examine the practical consequences of this decision, an exploratory check using the Holm correction was conducted and showed reduced sensitivity in several knowledge areas and for both biases. Finally, the sensitivity analysis is conditional on the criteria selected in Step 3 and is interpreted as an assessment of robustness and internal calibration, not as a set of independent confirmatory tests.

Beyond this potential extension, the model does present certain limitations. First, PROMSET currently addresses a single bias per iteration. The cross-bias analysis confirmed this limitation, as mitigating GLB did not mitigate GB, and vice versa. Thus, scenarios involving multiple simultaneous biases with heterogeneous impacts would require either sequential mitigation cycles or development of a multi-bias version. Second, in its current formulation, PROMSET is based on two related preference functions: the V-shape function and the linear function with an indifference threshold adjusted to mitigate bias. This structure allows for a gradual preference to be maintained and introduces an indifference zone only when the mitigation mechanism is activated; therefore, the current formulation is suitable for continuous quantitative criteria. Its direct application to qualitative criteria or scales with few levels is, therefore, limited and would require adapting the mitigation mechanism to other preference function structures. Third, both the identification of affected criteria and the calibration of preference thresholds depend on statistical choices and data quality. In Step 3, the criterion-level significance filter was applied without a multiple-comparisons correction, which increases sensitivity but also entails a greater risk of identifying some unaffected criteria as biased. Threshold calibration, in turn, depends on the robustness of the bias estimates, so sampling variability or imprecise estimates may propagate into the adjusted preference functions. The empirical relationship represented by Eq. (20) is specific to the institutional dataset analyzed here and should not be interpreted as a directly transferable calibration rule. What can be transferred to other applications is the sensitivity-analysis procedure used to derive such a relationship. Fourth, the mitigation mechanism treats the estimated association between average course grade and SET scores as bias. Meta-analytic evidence indicates that the relationship between SET ratings and student learning is small or statistically non-significant [21,62]. However, because no independent measure of student learning was available in the institutional data, it was not possible to determine whether part of the grade-related association reflected genuine learning rather than grading leniency. If so, the mitigation could remove part of a legitimate learning-related signal along with the bias. Finally, statistical non-significance should not be interpreted as evidence of a true null effect. Although the area-specific analysis helps isolate the influence of knowledge area, it may provide less statistical power to detect small residual bias effects than a pooled analysis.

7. Conclusions

This study introduced PROMSET, a structural extension of PROMETHEE II designed to mitigate systematic non-instructional biases in survey-based multicriteria evaluation systems. It achieves this by integrating robust MLR for bias identification with a comparison-dependent activation mechanism embedded within the preference functions for multicriteria aggregation.

Empirical validation based on SET data from 45,357 surveys covering 879 courses across four academic areas confirmed the method's robustness under heterogeneous quantitative conditions, encompassing both continuous and categorical bias variables. In addition to controlling evaluations by academic area, two of the most prominent non-instructional biases identified in the SET literature were mitigated: grade leniency and instructor gender. Furthermore, the model substantially improved the discriminative capacity of the ranking, reducing technical ties. These results demonstrate that bias mitigation arises from the structural properties of the proposed outranking mechanism rather than from threshold parametrization alone. By assessing 12 criteria grouped into four dimensions, PROMSET identified specific strengths of instructors, facilitating more accurate academic decision-making and tailored improvement plans.

Beyond the educational context, the proposed method contributes to the multicriteria decision-making literature by formalizing a bias-sensitive extension of outranking aggregation applicable to survey-based evaluation systems. PROMSET improves the fairness and transparency of faculty evaluation processes by providing more reliable and bias-mitigated assessments, thereby supporting evidence-based decision-making in hiring, promotion, tenure, and professional development. The ability to accurately identify teaching strengths and weaknesses also enables institutions to design targeted training initiatives and personalized support strategies, ultimately improving the quality of teaching and student learning outcomes. Furthermore, by increasing confidence in assessment results, PROMSET can help foster greater trust and participation among teachers in assessment systems, thereby promoting the development of a more constructive institutional culture around teaching excellence.

The proposed method is both scalable and transferable, eliciting criterion weights once per evaluation cycle while deriving bias-mitigation parameters from empirical data. Threshold calibration is then supported by a sensitivity-analysis procedure that can be replicated in other settings to derive context-specific empirical relationships for parameter adjustment. Moreover, the method's successful application across various disciplines and bias types further supports its versatility. This integration of econometric bias detection with outranking aggregation

Table A.1
Description of criteria.

Dimension	Criterion	SET question
Instructor's Personality (IP)	Punctuality	To what extent has this professor been punctual?
	Respect	Does the professor interact with students in a close and respectful manner?
	Accessibility	Is the professor accessible and willing to help students?
Instructor's Knowledge and Teaching Abilities (IKT)	Mastery	Does the professor have a strong command of this subject?
	Clarity	Is the professor clear in their explanations?
	Learning	Do I feel that I am learning with this professor?
Course Attributes (CA)	Planning	Does the professor properly distribute the course topics throughout the semester?
	Relevance	Do the activities proposed by the professor, both inside and outside the classroom, help achieve the course objectives?
	Resources	Does the professor use appropriate materials and resources (cases, exercises, readings, bibliography) for the proposed activities?
Assessment and Grading (AG)	Feedback	Does the professor provide students with personalized and useful feedback on their progress?
	Transparency	Has the professor clearly explained the course schedule (objectives, content, methodology, evaluation systems, etc.)?
	Coherence	Is there coherence between the course objectives and the evaluation systems?

opens new avenues for incorporating statistical correction mechanisms directly into multicriteria ranking procedures.

Finally, future research could explore integrating PROMSET's bias mitigation mechanism with classification methods, which may represent a relevant extension within multicriteria decision-making approaches to produce interpretable performance categories. Further studies could also extend the empirical validation of PROMSET by applying the model to datasets from non-European universities, thereby assessing its generalizability across different cultural and institutional environments. Future research could also explore prioritization mechanisms that jointly consider statistical evidence and effect magnitude when several candidate biases are considered for mitigation. Moreover, it would be valuable to develop extensions that enable the simultaneous mitigation of multiple biases. A promising direction is to recast this challenge as a group decision-making problem, where different institutional stakeholders each define a PROMSET specification targeting the bias they consider most critical, with parameter values guided by a sensitivity analysis analogous to that in Section 5.3, and aggregate the resulting rankings using established PROMETHEE-based group decision support procedures [61]. This would allow multiple biases to be addressed in parallel while embedding threshold selection within a structured institutional deliberation process.

Declaration of the use of AI

During the preparation of this manuscript, the authors used ChatGPT solely to improve the clarity and readability of the text. The authors reviewed and edited all generated content as necessary and take full responsibility for the final version of the manuscript.

Funding

This work was supported by the Santalucía Chair of Analytics for Education at Universidad Pontificia Comillas, the RENDIQUANT project (Ref. AVANZA_2025_2, Universidad Pontificia Comillas, Facultad de Ciencias Económicas y Empresariales), and RENDISTEM project (Ref. PR27/25-32464, Universidad Complutense de Madrid).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A

See Table A.1.

Table B.1
Evaluation criteria table for GLB mitigation in the area of quantitative methods.

	Criteria											
	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}	g_{11}	g_{12}
T	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5
T^*	1.5	1.5	1.5	1.5	3.0	3.0	3.0	3.0	3.0	1.5	3.0	3.0
q^*	0.0	0.0	0.0	0.0	0.7	0.6	0.4	0.5	0.4	0.0	0.4	0.4

Table B.2
Evaluation criteria table for GLB mitigation in the area of economics.

	Criteria											
	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}	g_{11}	g_{12}
T	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5
T^*	1.5	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0
q^*	0.0	0.7	0.8	0.5	0.9	1.0	0.8	0.9	0.9	1.1	0.8	0.7

Table B.3
Evaluation criteria table for GLB mitigation in the area of finance.

	Criteria											
	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}	g_{11}	g_{12}
T	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5
T^*	1.5	3.0	3.0	1.5	1.5	3.0	1.5	3.0	1.5	3.0	1.5	1.5
q^*	0.0	0.4	0.4	0.0	0.0	0.5	0.0	0.4	0.0	0.5	0.0	0.0

Table B.4
Evaluation criteria table for GLB mitigation in the area of international relations.

	Criteria											
	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}	g_{11}	g_{12}
T	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5
T^*	1.5	3.2	3.2	1.5	3.2	3.2	3.2	3.2	3.2	3.2	3.2	3.2
q^*	0.0	1.0	1.0	0.0	0.6	0.7	0.7	0.8	0.8	1.4	0.8	0.7

Appendix B

This appendix presents the parameters used in the mitigation of grade leniency bias (GLB) for each area of knowledge analyzed, based on a common set of evaluation criteria: Punctuality (g_1), Respect (g_2), Accessibility (g_3), Mastery (g_4), Clarity (g_5), Learning (g_6), Planning (g_7), Relevance (g_8), Resources (g_9), Feedback (g_{10}), Transparency (g_{11}), and Coherence (g_{12}). In all cases, each criterion was assigned an equal weight of 1/12 (8.33%), all criteria were defined as maximization criteria, and the indifference threshold (q) was set to zero. The bias-mitigated threshold (q^*) is applied only to criteria identified as biased (see Tables B.1–B.4).

Data availability

The authors do not have permission to share data.

References

[1] J.-F. Chen, H.-N. Hsieh, Q.H. Do, Evaluating teaching performance based on fuzzy AHP and comprehensive evaluation approach, Appl. Soft Comput. 28 (2015) 100–108, http://dx.doi.org/10.1016/j.asoc.2014.11.050.

[2] D. Santín, J. Tejada, Is it worth it? Using DEA to analyze the efficiency gains and costs of merging university departments: a case study of the complutense university of madrid, Int. Trans. Oper. Res. 32 (2025) 2593–2619, http://dx.doi.org/10.1111/itor.13488.

[3] S. Cunningham-Nelson, M. Baktashmotlagh, W. Boles, Visualizing student opinion through text analysis, IEEE Trans. Educ. 62 (2019) 305–311, http://dx.doi.org/10.1109/TE.2019.2924385.

[4] R.J. Kreitzer, J. Sweet-Cushman, Evaluating student evaluations of teaching: a review of measurement and equity bias in SETs and recommendations for ethical reform, J. Acad. Ethics 20 (2022) 73–84, http://dx.doi.org/10.1007/s10805-021-09400-w.

[5] A. Vanacore, M.S. Pellegrino, How reliable are students' evaluations of teaching (SETs)? A study to test student's reproducibility and repeatability, Soc. Indic. Res. 146 (2019) 77–89, http://dx.doi.org/10.1007/s11205-018-02055-y.

[6] S. Cook, D. Watson, R. Webb, Performance evaluation in teaching: dissecting student evaluations in higher education, Stud. Educ. Eval 81 (2024) 101342, http://dx.doi.org/10.1016/j.stueduc.2024.101342.

[7] R.O. Gilbert, D.R. Gilbert, Student evaluations of teaching do not reflect student learning: an observational study, BMC Med. Educ. 25 (2025) 313, http://dx.doi.org/10.1186/s12909-025-06896-3.

[8] W. Stroebe, Student evaluations of teaching encourages poor teaching and contributes to grade inflation: a theoretical and empirical analysis, Basic Appl. Soc. Psychol. 42 (2020) 276–294, http://dx.doi.org/10.1080/01973533.2020.1756817.

[9] C. Chisadza, N. Nicholls, E. Yitbarek, Race and gender biases in student evaluations of teachers, Econ. Lett. 179 (2019) 66–71, http://dx.doi.org/10.1016/j.econlet.2019.03.022.

[10] S. Siddiqui, C. Lento, Exploring teaching effectiveness and research on teaching and learning at AACSB accredited business schools in Canada and the US, Int. J. Educ. Manag. 36 (2022) 576–592, http://dx.doi.org/10.1108/IJEM-12-2021-0462.

[11] T.A. Mohammed, S.M. Pandhiani, Analysis of factors affecting student evaluation of teaching effectiveness in saudi higher education: the case of Jubail University College, Am. J. Educ. Res. 5 (2017) 464–475, http://dx.doi.org/10.12691/education-5-5-2.

[12] A. Ishizaka, P. Nemery, Multi-Criteria Decision Analysis: Methods and Software, Wiley, Chichester, 2013, http://dx.doi.org/10.1002/9781118644898.

[13] Y. Hadad, B. Keren, G. Naveh, The relative importance of teaching evaluation criteria from the points of view of students and faculty, Assess. Eval. High. Educ. 45 (2020) 447–459, http://dx.doi.org/10.1080/02602938.2019.1665623.

[14] B. Uttl, D. Smibert, Student evaluations of teaching: teaching quantitative courses can be hazardous to one's career, PeerJ 5 (2017) e3299, http://dx.doi.org/10.7717/peerj.3299.

[15] N. Radchenko, Student evaluations of teaching: unidimensionality, Subj. Biases, Educ. Econ. 28 (2020) 549–566, http://dx.doi.org/10.1080/09645292.2020.1814997.

[16] G. Wang, A. Williamson, Course evaluation scores: valid measures for teaching effectiveness or rewards for lenient grading? Teach. High. Educ. 27 (2022) 297–318, http://dx.doi.org/10.1080/13562517.2020.1722992.

[17] K. Chávez, K.M. Mitchell, Exploring bias in student evaluations: gender, race, and ethnicity, PS Polit. Sci. Polit. 53 (2020) 270–274, http://dx.doi.org/10.1017/S1049096519001744.

[18] F. Mengel, J. Sauermann, U. Zölitz, Gender bias in teaching evaluations, J. Eur. Econ. Assoc. 17 (2019) 535–566, http://dx.doi.org/10.1093/jeea/jvx057.

[19] T. Heffernan, Sexism, racism, prejudice, and bias: a literature review and synthesis of research surrounding student evaluations of courses and teaching, Assess. Eval. High. Educ. 47 (2022) 144–154, http://dx.doi.org/10.1080/02602938.2021.1888075.

[20] Z. Berezvai, G.D. Lukáts, R. Molontay, Can professors buy better evaluation with lenient grading? The effect of grade inflation on student evaluation of teaching, Assess. Eval. High. Educ. 46 (2021) 793–808, http://dx.doi.org/10.1080/02602938.2020.1821866.

[21] B. Uttl, C.A. White, D.W. Gonzalez, Meta-analysis of faculty's teaching effectiveness: student evaluation of teaching ratings and student learning are not related, Stud. Educ. Eval. 54 (2017) 22–42, http://dx.doi.org/10.1016/j.stueduc.2016.08.007.

[22] J.L. Arroyo-Barrigüete, A. Obregón, J.M. Ortiz-Lozano, A. Rua-Vieites, Spain Is Not Different: Teaching Quantitative Courses Can Also Be Hazardous To One's Career (At Least in Undergraduate Courses), PeerJ, 2022, http://dx.doi.org/10.7717/peerj.13456.

[23] N. Wagner, M. Rieger, K. Voorvelt, Gender, ethnicity and teaching evaluations: evidence from mixed teaching teams, Econ. Educ. Rev. 54 (2016) 79–94, http://dx.doi.org/10.1016/j.econedurev.2016.06.004.

[24] P. Spooren, B. Brockx, D. Mortelmans, On the validity of student evaluation of teaching: the state of the art, Rev. Educ. Res. 83 (2013) 598–642, http://dx.doi.org/10.3102/0034654313496870.

[25] L. Zhao, P. Xu, Y. Chen, S. Yan, A literature review of the research on students' evaluation of teaching in higher education, Front. Psychol 13 (2022) 1004487, http://dx.doi.org/10.3389/fpsyg.2022.1004487.

[26] A. Boring, K. Ottoboni, P.B. Stark, Student evaluations of teaching (mostly) do not measure teaching effectiveness, Sci. Res. (2016) 1–11, http://dx.doi.org/10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1.

[27] D.A.M. Peterson, L.A. Biederman, D. Andersen, T.M. Ditonto, K. Roe, Mitigating gender bias in student evaluations of teaching, PLoS One 14 (2019) e0216241, http://dx.doi.org/10.1371/journal.pone.0216241.

[28] A. Boring, A. Philippe, Reducing discrimination in the field: evidence from an awareness raising intervention targeting gender biases in student evaluations of teaching, J. Public Econ. 193 (2021) 104323, http://dx.doi.org/10.1016/j.jpubeco.2020.104323.

[29] F.F. Kim, L.A. Williams, E.L. Johnston, Y. Fan, Bias intervention messaging in student evaluations of teaching: the role of gendered perceptions of bias, Heliyon 10 (2024) e37140, http://dx.doi.org/10.1016/j.heliyon.2024.e37140.

[30] J.L. Arroyo-Barrigüete, M.C. Bada Olaran, L.I. Lazcano Benito, J. Márquez Vigil, J.M. Ortiz Lozano, A. Rúa Vieites, Is it possible to redress noninstructional biases in student evaluation of teaching surveys? Quantitative analysis in accounting and finance courses, Stud. Educ. Eval 77 (2023) 101263, http://dx.doi.org/10.1016/j.stueduc.2023.101263.

[31] N. Zipser, L. Mincieli, A framework for using SET when evaluating faculty, Stud. High. Educ. 50 (2025) 168–182, http://dx.doi.org/10.1080/03075079.2024.2332412.

[32] Y. Chen, P.-C. Chen, Using a novel conjunctive MCDM approach based on DEMATEL, fuzzy ANP, and TOPSIS as an innovation support system for Taiwanese higher education, Expert. Syst. Appl. 37 (2010) 1981–1990, http://dx.doi.org/10.1016/j.eswa.2009.06.079.

[33] F.H. Abanda, E.L. Chia, K.E. Enongene, M.B. Manjia, K. Fobissie, U.J.M.N. Pettang, C. Pettang, A systematic review of the application of multi-criteria decision-making in evaluating nationally determined contribution projects, Decis. Anal. J 5 (2022) 100140, http://dx.doi.org/10.1016/j.dajour.2022.100140.

[34] C. Giannoulis, A. Ishizaka, A web-based decision support system with ELECTRE III for a personalised ranking of British universities, Decis. Support Syst. 48 (2010) 488–497, http://dx.doi.org/10.1016/j.dss.2009.06.008.

[35] J.M. Conejero, J.C. Preciado, A.E. Prieto, M.C. Bas, V.J. Bolós, Applying data driven decision making to rank vocational and educational training programs with TOPSIS, Decis. Support Syst. 142 (2021) 113470, http://dx.doi.org/10.1016/j.dss.2020.113470.

[36] S. Kiris, AHP and multichoice goal programming integration for course planning, Int. Trans. Oper. Res. 21 (2014) 819–833, http://dx.doi.org/10.1111/itor.12081.

[37] D. Deliktas, O. Ustun, Student selection and assignment methodology based on fuzzy MULTIMOORA and multichoice goal programming, Int. Trans. Oper. Res. 24 (2017) 1173–1195, http://dx.doi.org/10.1111/itor.12185.

[38] J.W. Gong, H.C. Liu, X.Y. You, L. Yin, An integrated multi-criteria decision making approach with linguistic hesitant fuzzy sets for E-learning website evaluation and selection, Appl. Soft Comput. 102 (2021) 107118, http://dx.doi.org/10.1016/j.asoc.2021.107118.

[39] S.L. Xu, T. Yeyao, M. Shabaz, Multi-criteria decision making for determining best teaching method using fuzzy analytical hierarchy process, Soft Comput. 27 (2023) 2795–2807, http://dx.doi.org/10.1007/s00500-022-07554-2.

[40] R.H.C. Machado, S.V. Conceição, R. Pelissari, S. Ben Amor, T.L. Resende, A multiple criteria framework to assess learning methodologies, Think. Ski. Creat 48 (2023) 101290, http://dx.doi.org/10.1016/j.tsc.2023.101290.

[41] Y. Qin, S.R.M. Hashim, J. Sulaiman, An interval AHP technique for classroom teaching quality evaluation, Educ. Sci 12 (736) (2022) http://dx.doi.org/10.3390/educsci12110736.

[42] Q.H. Do, V.T. Tran, T.T. Tran, Evaluating lecturer performance in Vietnam: an application of fuzzy AHP and fuzzy TOPSIS methods, Heliyon 10 (e30772) (2024) http://dx.doi.org/10.1016/j.heliyon.2024.e30772.

[43] M. Popović, G. Savić, M. Kuzmanović, M. Martić, Using data envelopment analysis and multi-criteria decision-making methods to evaluate teacher performance in higher education, Symmetry 12 (563) (2020) http://dx.doi.org/10.3390/sym12040563.

[44] Z. Chen, W. Liang, S. Luo, A novel integrated picture fuzzy MACONT method and its application in teaching quality evaluation in higher education, IEEE Access 12 (2024) 88345–88356, http://dx.doi.org/10.1109/ACCESS.2024.3418035.

[45] S. Chakraborty, TOPSIS and modified TOPSIS: A comparative analysis, Decis. Anal. J 2 (2022) 100021, http://dx.doi.org/10.1016/j.dajour.2021.100021.

[46] U. Afzal, M.R. Ahmad, N. Alessa, N. Raza, F.M.O. Birkea, S. Alkhalaf, N. Omer, Intelligent faculty evaluation and ranking system based on N-framed plithogenic fuzzy hypersoft set and extended NR-TOPSIS, Alex. Eng. J. 109 (2024) 18–28, http://dx.doi.org/10.1016/j.aej.2024.08.071.

[47] M. Radovanović, D. Milinković, Ž. Stević, D. Pamučar, Evaluation of university professors using spherical fuzzy AHP and grey MARCOS method, Decis. Mak. Appl. Manag. Eng. 8 (2025) 192–210, http://dx.doi.org/10.31181/sdmap21202518.

[48] K.R. Mohan, P.S. Reddy, G.S. Prasad, MCDM PROMETHEE method in identifying best teacher awardee, in: R.K. Sharma, M.K. Tiwari, P.K. Kapur (Eds.), Decision Making under Uncertainty via Optimization, Modelling, and Analysis, Springer, Singapore, 2025, pp. 503–514, http://dx.doi.org/10.1007/978-981-96-0085-4_26.

[49] E. Thanassoulis, P.K. Dey, K. Petridis, I. Goniadis, A.C. Georgiou, Evaluating higher education teaching performance using combined analytic hierarchy process and data envelopment analysis, J. Oper. Res. Soc. 68 (2017) 431–445, http://dx.doi.org/10.1057/s41274-016-0165-4.

[50] M.D. Bailey, D. Michaels, An optimization based DSS for student to teacher assignment: classroom heterogeneity and teacher performance measures, Decis. Support Syst. 119 (2019) 60–71, http://dx.doi.org/10.1016/j.dss.2019.02.006.

[51] G. Ottomano Palmisano, R. Sardaro, P. La Sala, Recovery and resilience of the inner areas: Identifying collective policy actions through PROMETHEE II, Land 11 (2022) 1181, http://dx.doi.org/10.3390/land11081181.

[52] M.M. Silva, A.P.H. de Gusmão, A.P.C.S. Costa, An outranking multicriteria method for nominal classification problems with minimum performance profiles, Math. Probl. Eng. 2019 (2019) 4078909, http://dx.doi.org/10.1155/2019/4078909.

[53] J. Rosenfeld, Y. De Smet, An extension of PROMETHEE to hierarchical multicriteria clustering, Int. J. Multicriteria Decis. Mak. 8 (2020) 133–150, http://dx.doi.org/10.1504/IJMCDM.2019.106911.

[54] J.P. Brans, Y. De Smet, PROMETHEE methods, in: S. Greco, M. Ehrgott, J. Figueira (Eds.), Multiple Criteria Decision Analysis: State of the Art Surveys, Vol. 1, second ed., Springer, New York, 2016, pp. 187–219.

[55] T.L. Saaty, K. Peniwati, Group Decision Making: Drawing Out and Reconciling Differences, RWS Publications, Pittsburgh, PA, 2008.

[56] M. Braga, M. Paccagnella, M. Pellizzari, Evaluating students' evaluations of professors, Econ. Educ. Rev. 41 (2014) 71–88, http://dx.doi.org/10.1016/j.econedurev.2014.04.002.

[57] O.K. Baruch, The persistence of gender bias in student evaluations of teaching: the role of gender stereotypes, J. Acad. Ethics 23 (2025) 279–303, http://dx.doi.org/10.1007/s10805-024-09535-6.

[58] D.V. Chulkov, J. Van Alstine, Challenges in designing student teaching evaluations in a business program, Int. J. Educ. Manag. 26 (2012) 162–174, http://dx.doi.org/10.1108/09513541211201979.

[59] M.G. Kendall, A new measure of rank correlation, Biometrika 30 (1938) 81–93, http://dx.doi.org/10.2307/2332226.

[60] F. Barrera, M. Segura, C. Maroto, Multiple criteria decision support system for customer segmentation using a sorting outranking method, Expert Syst. Appl. 238 (2024) 122310, http://dx.doi.org/10.1016/j.eswa.2023.122310.

[61] C. Macharis, J.P. Brans, B. Mareschal, The GDSS PROMETHEE procedure: a PROMETHEE-GAIA based procedure for group decision support, J. Decis. Syst. 7 (1998) 283–307.

[62] D.E. Clayson, Student evaluations of teaching: are they related to what students learn? A meta-analysis and review of the literature, J. Mark. Educ. 31 (2009) 16–30, http://dx.doi.org/10.1177/0273475308324086.