On the impact of weighting schemes on alternatives' evaluation

Pavel Novoa-Hernández[0000−0003−3267−6753], David A. Pelta[0000−0002−7653−1452], and José Luis Verdegay[0000−0003−2487−942X]

Department of Computer Science and A.I.
Universidad de Granada, 18014, Granada , Spain
{pavelnovoa,dpelta,verdegay}@ugr.es

Abstract. Performance evaluation is ubiquitous nowadays. A fundamental approach to assessing alternatives (e.g., students, employees, universities, countries) involves weighting criteria to calculate their final scores. Determining these weights is crucial, especially when alternatives are evaluated individually and without comparison with others. In such cases, achieving consensus on weights among evaluators is vital for fairness. In the context of students' evaluation from a university course, this initial study analyzes whether different weight assignments significantly alter final scores and identifies which of those assignments lead to the highest scores. The data consists of 53 students across seven assignments, and a bootstrap analysis is performed for validation. Results indicate statistically significant score variations compared to predetermined weights, with the 'Sum of Ranks' approach resulting in the highest scores over 90% of the time. This highlights the importance of weighting schemes and recommends the use of bootstrapping to justify their selection.

Keywords: weights determination · flexible criteria preference · ranked weights · uncertainty

1 Introduction

Performance evaluation is ubiquitous nowadays. From employees to students, from universities to countries, and so on, many things are measured and assessed with different objectives in mind. In the most basic setting, the requirement is to assign a score to an entity or alternative a from a set of numerical criteria c_i. For that purpose, an evaluation or scoring function should be provided. The most frequently used function is a weighted linear aggregation rule.

Key aspects are how the weights are determined and which is their impact on the final scores. Despite this, often the determination of weights is generally considered a minor problem: a "decision maker" defines a specific weighting scheme using his/her knowledge, considering that if the criteria c_i is more important than the criteria c_j, then the weights should satisfy $w_i > w_j$.

An overlooked aspect here and a clear source of uncertainty is that there are potentially infinite sets of weights that can be used. And each specific set of weights impacts the final score that an alternative can achieve. This latter aspect is not quite relevant if the aim is to produce a ranking for a set of entities. In such a case, just a relation among the scores is needed (without taking care of their magnitudes). In other words, it is said that an alternative A is better than B if $score(A) > score(B)$.

Here we are interested in the case where the final score of an alternative is relevant *per se* and not in relation with those from other ones (as it happens in multicriteria decision-making problems).

Focusing on the context of the academic evaluation of students from a Computer Programming Methodology course at the University of Granada, this contribution aims to provide insights into the following research questions:

- **Q1:** Does the use of different weight schemes produce significantly different distributions of the final scores?
- **Q2:** Is there any set of weights (among the tested ones) that consistently produces the highest final scores?

The rest of the work is organized as follows. Section 2 presents the basic settings and the weights' approximation methods. The case study, which provides the initial insights for the questions posed before, appears in Section 3. Then, in Section 4 a bootstrap analysis is performed to confirm or not the findings in the case study. Finally, the conclusions and further discussions are provided in Section 5.

2 Problem description

As we stated before, the starting point is a set of m alternatives $a_1, a_2, \ldots, a_m$ that are evaluated over a set of n criteria $c_1, c_2, \ldots, c_n$. This information is organized in a matrix, $E_{m\times b}$ where each entry $e_{ij} \in \mathbb{R}$ corresponds to the score obtained by the alternative a_i in the c_j criterion.

In the most basic setting, the requirement is to assign a score S to an alternative $a_i, i \in 1, \ldots, m$ from the set of numerical criteria. One of the most frequently used functions to calculate such a score is a weighted linear aggregation rule:

$$S(a_i) = \sum_{j=1}^{n} w_j \times e_{ij} \tag{1}$$

$$w_j \in [0, 1] \tag{2}$$

$$w_1 + w_2 + \ldots + w_n = 1 \tag{3}$$

If $w_i > w_j$ then the criterion c_i is more important than c_j.

It should be noted that in our particular case, as the values $e_{ij} \in [0, 10]$, then no normalization is required.

When several evaluators are involved in the scoring process, reaching a consensus about the weights may be far from trivial. In turn, we consider that those evaluators can agree on the order of importance of the criteria.

Suppose that three criteria are available. It is simpler to agree in "sorting" them in terms of importance as $c_2 \succeq_p c_1 \succeq_p c_3$ (which should be translated numerically to order between the weights $w_2 > w_1 > w_3$, satisfying conditions 2, 3), than to agree in a vector of weights like $(0.3, 0, 5, 0.2)$ or $(0.35, 0.4, 0.25)$.

An overlooked aspect in the determination of such specific weights, and a clear source of uncertainty, is that there are potentially infinite sets of weights that can be used. As a consequence, the use of a specific set of weights impacts the final score that an alternative can achieve [9,10]. This situation is discussed in the next subsection.

2.1 Weights determination

The question "How to determine the weights?" has different answers, depending on the information available regarding the "true" weights. Three situations are identified in [8]: *"knowing nothing about the true weights, knowing the rank order information, and knowing the relative weight information"*.

As we are in a situation where only rank order information can be captured, we focus on the so-called Weights Approximation Methods [4] or ranked weights. These methods, starting from a decreasing ordering of the weights (in terms of importance) propose formulas for determining their specific values, given the constraints defined before.

Specifically, three methods are considered: rank order centroid (ROC) [3], the sum of ranks (RS) [12] and the reciprocal of the ranks (RR) [12].

Assuming $w_1 > w_2 > \ldots > w_n$, these sets of weights are defined as follows:

$$\text{Rank Order Centroid } \textbf{(ROC):}\ w_j = \frac{1}{n}\sum_{k=j}^{n}\frac{1}{k}$$

$$\text{Sum of Ranks } \textbf{(RS):}\ w_j = \frac{n+1-j}{\sum_{k=1}^{n} k} = \frac{2(n+1-j)}{n(n+1)}$$

$$\text{Rank Reciprocal } \textbf{(RR):}\ w_j = \frac{1/j}{\sum_{k=1}^{n} 1/k}$$

3 Case study: impact of weights in students evaluation

The evaluation of a student's performance is one of the many tasks that a teacher must do. How to make such an evaluation has been extensively explored in various settings, encompassing the examination of the elements that account for it [1,6,11] and the development of practical measuring tools [2].

The problem can be approached from different perspectives. For example, two multi-criteria (MD) decision approaches were used in [15] to evaluate student performance: simple multi-attribute assessment (SMART [13]) and multi-objective optimization of proportions analysis (MOORA [5]). The authors employed methods based on entropy and utility calculations to determine the weights of the criteria and sub-criteria. Other authors focused on the impact of different sets of weights. For example, in a recent work [7], students were asked to select with which weighting scheme want to be evaluated. The conclusion was that "the majority of students in both courses examined in the current study did not select weighting schemes that resulted in their highest potential grade". In [14], the authors explored the impact of different weights in the TOPSIS method and concluded that "the selection of an adequate weighting method has a significant impact on the overall results of the TOPSIS technique."

To address the research questions, we resort to information available from a specific group of 53 students from a Computer Programming Methodology subject who completed seven lab assignments. This example has several relevant features: students are assessed individually, the final score is relevant *per se*, more than one evaluator is available, and the students are required to know both the scoring function and the weighting scheme at the beginning of the year.

The assignments, which include programming exercises, software project development, presentations, etc., are called $E_1, E_2, \ldots, E_7$ in what follows. Due to the incremental nature of the topics considered in the subject, instructors agree that the latter assignments are more important than the initial ones. After a meeting, the following order of importance was agreed:

$$E_7 \succeq_p E_6 \succeq_p E_5 \succeq_p E_4 \succeq_p E_3 \succeq_p E_2 \succeq_p E_1 \tag{4}$$

This order implies the following relation for the weights

$$w_7 > w_6 > w_5 > w_4 > w_3 > w_2 > w_1 \tag{5}$$

where w_j is the weight for the E_j assignment.

Figure 1(left) displays a table with the sets of weights calculated with the weights approximation methods considered. The column G-W contains a set of weights used by a specific instructor in a previous year. Figure 1(right) shows a visualization of the different weight assignments produced by each method. A clear difference is immediately observed for the values of weight w_7 assigned to the most important assignment E_7. Regarding its contribution to the final score, w_7 represents from 25% in the RS scheme, up to 38% in RR. In G-W the teacher assigns it a 30%.

If we consider the contribution of the last two assignments, we observe that $w_6 + w_7$ accounts for the 50% of the final score in G-W, around 60% in ROC and RR, and 46% in RS.

The final score for every student was calculated using every set of weights, and it is a real number between $[0 \ldots 10]$. As it is clear, every student has four candidate final grades (one per each set of weights), so we also calculate the average among them for the analysis. This value will be considered under the category AVG. For the first analysis, we translate the final scores into the usual linguistic values "*fail*" (score < 5), "*pass*" ($5 \leq$ score < 7), "*good*"($7 \leq$ score < 9) and "*excellent*" (score ≥ 9).

The question to answer here is: *Do the use of the different weights produce a different distribution in the number of students with "fail", "pass", "good" and "excellent" scores*?

	Weights			
Evaluation	G−W	ROC	RS	RR
E1	0.04	0.020	0.036	0.055
E2	0.06	0.044	0.071	0.064
E3	0.10	0.073	0.107	0.077
E4	0.14	0.109	0.143	0.096
E5	0.16	0.156	0.179	0.129
E6	0.20	0.228	0.214	0.193
E7	0.30	0.370	0.250	0.386

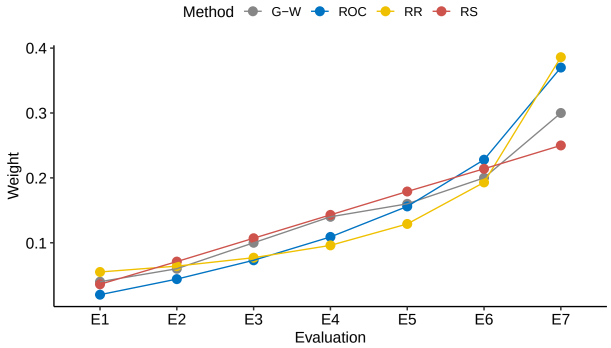

Fig. 1. Weights corresponding to the traditional scheme (G-W) and the ranked weights methods.

Figure 2 provides insights into the answer. The distribution of cases is different. In the case of "*fail*", the teacher weights (G-W), RS, and AVG make that 21% of the group does not reach the minimum to pass the lab assignments. For ROC and RR cases, the value goes up to 25%. The number of "*pass*" varies from 28% in RS, up to 43% in RR. For the G-W scheme, this value is 36%.

It is interesting to note that, in comparison with G-W, the RS scheme obtained the same values for "*fail*", a lower one for "*pass*" but higher ones for "*good*" and "*excellent*". The RS scheme allows achieving the highest rate of "*excellent*" evaluations (9%).

The RR scheme is the most "strict", leading to the highest percentage of *fail* and the lowest one for *good*+*excellent*.

Focusing now on the second question posed at the beginning, from the perspective of a student, he/she would like to be evaluated with the set of weights that produce the highest final score. So, we analyzed which set of weights allowed us to achieve the highest score for every student. Out of 53 cases, the highest values were achieved 44, 5, and 4 times by RS, ROC, and RR weights, respectively. The G-W scheme (the one used by the instructor) never allowed obtaining the highest evaluation.

One may ask if the increases in the scores are relevant. If we consider the differences between the scores produced by G-W and RS, the latter are, on average, 2.6% higher, with a standard deviation of 1.5. This may look like a minor improvement, but from the perspective of a student, it could be relevant.

4 Statistical assessment of the results

The evaluators agreed on the fact that the scores' distribution in the analyzed dataset, was similar to the ones they observed in their groups of students.

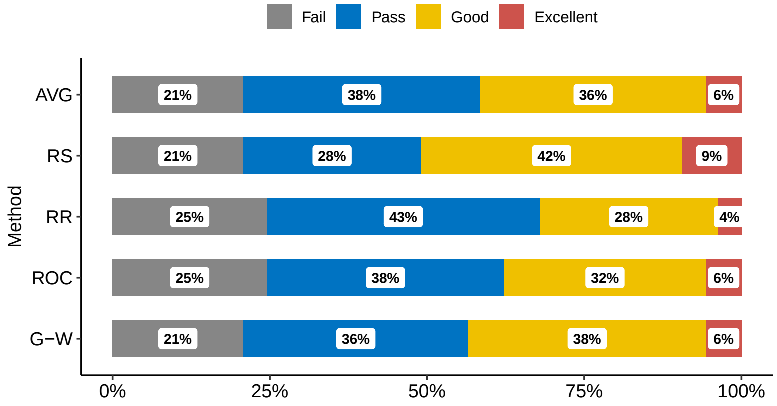

Fig. 2. Percentages of students in every scoring category according to the weight assignment methods.

So, we perform a bootstrap analysis for the estimation of the true statistics reported in the previous section and to allow more powerful statistical comparisons.

Departing from the previous set of 53 students, we performed a bootstrap sampling of 1000 replicates of size 50 with replacement. For each replicate, we aggregated 6 different measures for each set of weights. In addition to the proportion of *fail*, *pass*, *good*, and *excellent* students, we also measured the average score, and the proportion of times in which the given weight scheme allowed to attain the highest score for each student.

Figure 3 displays a set of box plots that summarizes the results obtained. The four plots on top (labeled as a, b, c, d), show the estimated percentage of students with a *fail, pass, good* and *excellent* scores as a function of the set of weights used. Plot (e) summarizes the results in terms of the final scores, while plot (f) shows the estimation of the number of times that each weight assignment will allow attaining the highest score.

To assess if the observed differences among the methods (weights assignments) are statistically significant (or not), we proceed as follows. First, a Friedman omnibus test to identify if there are significant differences among the methods was run. If such a difference exists, then a Wilcoxon *post-hoc* test with p-value adjustment by the Bonferroni method was performed to identify between which pair of methods such differences exist. Note that the results of the Friedman test are included at the top of each graph, while the results of the Wilcoxon test in method (x-axis) are marked with ****. In the latter case, the methods identified with that mark correspond to those significantly different from G-W.

Figure 3 shows that in all cases, Friedman's test returns a very low p-value (e.g. less than 0.001), indicating that the distribution of the cases produced by the weights assignment is different. Regarding the distribution of the score cate-

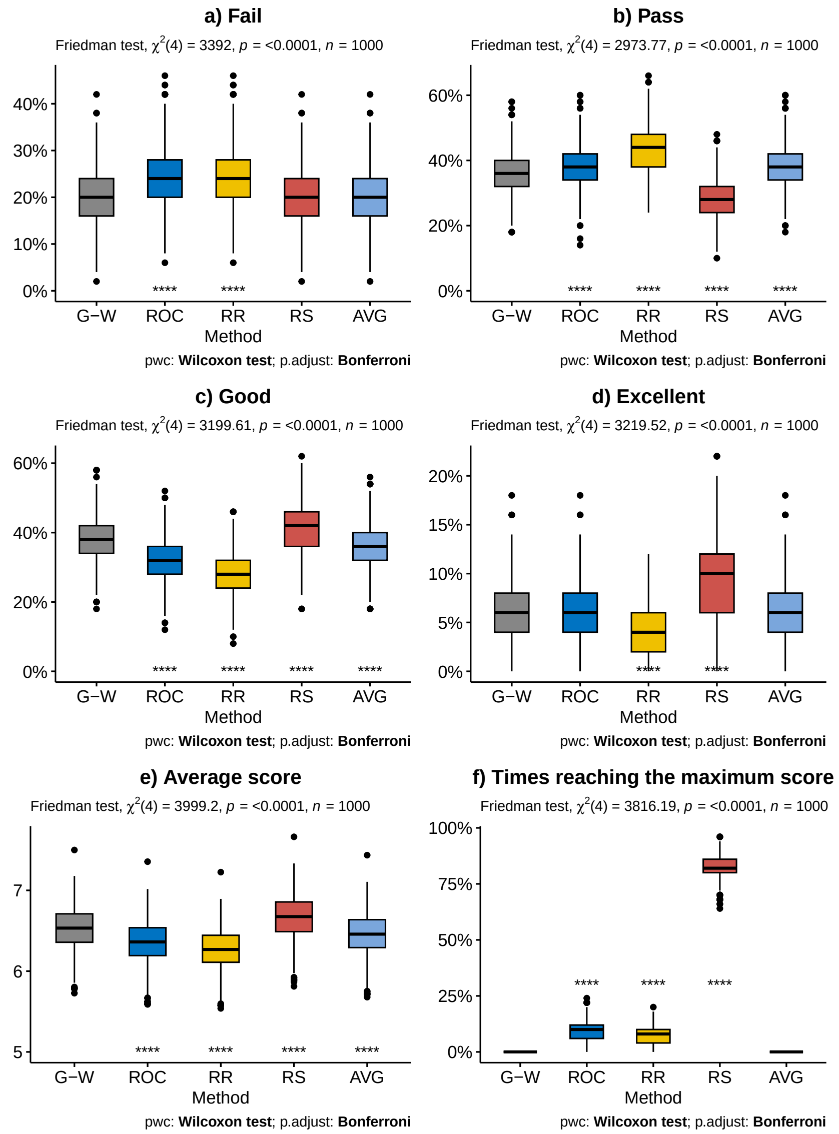

Fig. 3. Comparison of scores from different weight assignment schemes after a bootstrapping sampling of 1000 replicates. Weight assignments marked with **** indicate a statistical difference with the G-W approach.

gories (plots a-d from Fig.3) we observe that RR scores are significantly different from G-W in all cases. However, although the proportions of *fail* and *pass* are higher in favor of RR, they are lower for the *good* and *excellent* categories. In contrast, the RS method has the opposite behavior to RR. That is, it exhibits similar or lower proportions than G-W for *fail* and *pass*, but significantly higher for *good* and *excellent*.

If we consider the results in terms of the average score (Fig. 3-e) it becomes clear that all the weights assignment methods are significantly different from G-W. We can see that RS shows an overall distribution of scores significantly higher than G-W (Fig.3-e), while RR is significantly lower. Note that this is consistent with the previous analysis.

Finally, Fig.3-f shows that the RS scheme is the most beneficial method for students: in more than 60% of cases, using the weights calculated by RS allows them to attain the highest final grade. This difference is not only significant with G-W but also with the rest of the methods. The closest ones, ROC and RR, fail to exceed 30%.

5 Conclusions and further work

Performance evaluation of entities of different types is ubiquitous nowadays. Many times, the evaluation is made using a weighted sum approach (which is simple to understand and calculate).

In the context of the academic evaluation of university students and departing from the fact that the definition of weights is a key aspect, we wanted to shed some light on the following questions:

- **Q1:** Does the use of different weight schemes produce different distributions of the final scores?
- **Q2:** Is there any set of weights (among the tested ones) that consistently produces the highest final scores?

Regarding **Q1**, the answer is yes. The results confirmed that the use of different weights produced a different distribution of scores in terms of *fail*, *pass*, *good*, and *excellent* scores. When the bootstrap analysis was run, the differences were statistically significant.

Concerning the given set of weights (G-W), it was observed that some of the weighting schemes, like ROC, produced more *fails*. Other, like RS, produced a similar number of *fails*, but the students who obtained a $score >= 5$ were differently distributed among the *pass*, *good*, and *excellent* grades.

Regarding **Q2**, the answer is also yes. The set of weights derived from the "Sum of Ranks" approximation method allowed us to achieve (in most cases) the highest score for the students. In the case study, this fact happened to 44 out of 53 students.

The main difference between RS and the other weighting schemes is that in RS, the weights grow linearly as a function of the importance of the criteria. In contrast, in the rest of the schemes, this growth is low for weights of low importance, while very high for weights of higher importance. Such a behavior has important implications, at least for the proposed case study. In this context, the use of bootstrapping is a clear strategy to provide solid foundations for choosing a particular weighting scheme.

Now, from a practical point of view, the group of instructors agreed on using the RS scheme for future evaluations in the computer programming methodology subject. The initial impressions are very positive. Of course, there is here some room for discussion regarding why it is desired to obtain a high score consistently in student evaluations, or if different considerations such as bell-shaped scores from evaluations should be considered. In our experience, this is a highly controversial issue, and we prefer not to pose our opinion here.

From a research point of view, it is clear that some of the conclusions posed (for example, that "Sum of ranks" attained the highest score) cannot be directly generalized to other contexts. However, the research questions posed, and the analysis proposed can be readily extended to other problems with similar features (like the evaluation of the achievement of sustainable development goals or the evaluation of universities), considering both the bootstrap analysis and testing other weighting approximation methods.

Acknowledgments

The authors acknowledge support from project PID2020-112754GB-I00, MCIN/AEI 347/10.13039/501100011033, Ministry of Science and Innovation, Spain

References

1. Sarah A. Alwarthan, Nida Aslam, and Irfan Ullah Khan. Predicting student academic performance at higher education using data mining: A systematic review. *Applied Computational Intelligence and Soft Computing*, 2022, 2022.
2. Deborah L. Bandalos. *Measurement Theory and Applications for the Social Sciences*. Methodology in the Social Sciences. Guilford Publications, 2018.
3. F. Hutton Barron and Bruce E. Barrett. Decision quality using ranked attribute weights. *Management Science*, 42(11):1515–1523, nov 1996.
4. John Butler, Jianmin Jia, and James Dyer. Simulation techniques for the sensitivity analysis of multi-criteria decision models. *European Journal of Operational Research*, 103(3):531–546, 1997.
5. Shankar Chakraborty. Applications of the MOORA method for decision making in manufacturing environment. *The International Journal of Advanced Manufacturing Technology*, 54(9-12):1155–1166, oct 2010.
6. Jennifer Chung, Stephen McKenzie, Ashleigh Schweinsberg, and Matthew Edward Mundy. Correlates of academic performance in online higher education: A systematic review. *Frontiers in Education*, 7, 2022.
7. Paige Coyne and Sarah J. Woodruff. Giving students choice: Does the use of a flexible assessment weighting scheme result in better student grades? *International Journal of Teaching and Learning in Higher Education*, 33(3):398–406, 2022.
8. Dun Liu, Tianrui Li, and Decui Liang. An integrated approach towards modeling ranked weights. *Computers & Industrial Engineering*, 147:106629, 2020.
9. Pavel Novoa-Hernández, David Pelta, Carlos Cruz, and José Luis Verdegay. A multi-criteria flexible weighting approach for the academic assessment of university students *Conference on the University Teaching of Computer Science (JENUI)*, 2023 (In Spanish).

10. Pavel Novoa-Hernández, Boris Pérez-Cañedo, David A. Pelta, and José Luis Verdegay. Towards imprecise scores in multi-criteria decision making with ranked weights. In S. Massanet, S. Montes, D. Ruiz-Aguilera, and M. González-Hidalgo, editors, *Lecture Notes in Computer Science*, pages 197–207. Springer Nature Switzerland, 2023.
11. Carlos Felipe Rodríguez-Hernández, Eduardo Cascallar, and Eva Kyndt. Socioeconomic status and academic performance in higher education: A systematic review. *Educational Research Review*, 29, 2020.
12. William G. Stillwell, David A. Seaver, and Ward Edwards. A comparison of weight approximation techniques in multiattribute utility decision making. *Organizational Behavior and Human Performance*, 28(1):62–77, 1981.
13. James M. Taylor and Betty N. Love. Simple multi-attribute rating technique for renewable energy deployment decisions (SMART REDD). *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology*, 11(3):227–232, 2014.
14. Roman Vavrek. Evaluation of the impact of selected weighting methods on the results of the TOPSIS technique. *Int. J. Inf. Technol. Decis. Mak.*, 18(06):1821–1843, November 2019.
15. Masna Wati, Niken Novirasari, Edy Budiman, and Haeruddin. Multi-criteria decision-making for evaluation of student academic performance based on objective weights. In *2018 Third International Conference on Informatics and Computing (ICIC)*. IEEE, 2018.