ISSN 2712-0562
Sustainable Engineering and Innovation *Original Research*
Vol. 2, No. 2, December 2020, pp.89-101
https://doi.org/10.37868/sei.v2i2.112

How do weights affect faculty performance evaluations?

Özge Büyükdağlı[1]*, Sencer Yeralan[2]
[1]Computer Science and Engineering, International University of Sarajevo, Bosnia
[2] Agricultural and Biological Engineering, University of Florida, USA

*Corresponding author: obuyukdagli@ius.edu.ba

Abstract
Quite commonly, faculty performance evaluations use a weighted scheme. Individual faculty members are evaluated on a scale with respect to teaching, research, and service activities. These scores are then combined using pre-determined weights to obtain a combined score that is often used to compare different members. The presented study aimed to investigate the effects of selecting the weights on the individual scores and rankings. The interest is not on single faculty members, but rather on the systems aspects of the practice. That is, how do the weights affect the educational system as a whole? How sensitive is the evaluation system to the selection of the weights? In order to question the leverage, a decision-maker who determines the weights would have on the outcome of the rankings, the approach based on numerical examples and formal linear programming (LP) considerations is used.

Keywords: Faculty evaluation, Numerical evaluation schemes, Weights

1. Introduction

Quantitative faculty evaluations assign a performance score to each member. The widespread practice is to consider a set of criteria, to assign raw scores to the professors regarding each of the criteria, and then combine these raw scores into a weighted sum to obtain a single aggregate score. These scores are used by administration to various ends.

Faculty member evaluation is a challenging process due to the general nature of academic institutions. The literature on faculty evaluations is extensive and broad. There are many studies that discuss the need, appropriateness, fairness, and utility of this practice [11], [13], [17]. The evaluation results generally serve two main proposes: to improve faculty performance by assessing the strengths and weaknesses of the professors (formative purpose), and to help administrators to make personnel decisions like promotions or salary adjustments (summative purpose). The weight defined for each criterion reflects the strategic preferences of the institution. For example, teaching-oriented institutions may assign a larger weight to teaching than would research-oriented institutions. A general criticism is the quantification of what is essentially qualitative in nature [8], [12], [18]. Redmon [11] questions the conflicting objectives of this process: judging and assisting. Younes [3] also focuses on the issue of "competing values" and discusses the differences between the administrative perception and the faculty perception of the process. Gunn [2] claimed most evaluation systems are spurious and lead to rivalry among academics. So, there is a rising concern over effectiveness and credibility of evaluation systems from many different aspects.

The objectivity of an evaluation system may be determined by its consistency of conclusions based on the same data [16]. Beyond reliability and objectivity, several other questions emerge. For instance, would a

practice that uses the same set of weights for each faculty member promote duplications and reduce diversity? We note that there are some evaluation systems where customized weights are assigned at the beginning of each year specifically for each department or professor, according to the strategic directions and the expectations of the institution during that period [15].

There are formal models proposed for faculty performance evaluation which include the so-called Analytic Hierarchy Process approach [6], [9], [14] and Multi-Criteria Decision-Making techniques [1], [4], [7]. In particular, [4] propose a comprehensive model to design performance evaluation systems. In that study, the assignment of weights and their bounds are sought using formal arguments. However, as far as we know, the present work is the first to study the effects of weight assignments and the sensitivity of the evaluation system to these assignments in a formal setting.

Quite often, faculty responsibilities are a combination of three main duties: teaching, research, and service, each with its corresponding weight. Let these weights be represented by 3-tuples, such as (0.50, 0.30, 0.20), corresponding to a weight of 50% for teaching, 30% for research, and a weight of 20% for service. A web survey of many U.S. universities indicates that typical weights are relatively round numbers, such as (50%, 25%, 25%) or (40%, 40%, 20%), etc. Acknowledging the expediency of this practice, it is, nonetheless, open to criticism. Most view these three activities as being neither independent nor mutually exclusive [5], [19]. The fact that most weights are round numbers gives the impression that these weights are assigned based on rather ad hoc approximations or through deliberations, e.g. Delphi methods, rather than careful quantitative scrutiny of their effects.

Irrespectively, it is clear that aggregate scores have a significant impact on the institution, be it merit pay, strategic alignment, or personnel moral. This observation that weights seem to be assigned in a intuitive manner gives justification to our extensive formal quantitative work. We focus on the process of assigning weights to criteria. How significantly does the assignment of weights affect the resultant evaluation? How does the institution justify the assignment of weights? Or, how can one rigorously identify the causality relationship between weight assignments and the alignment of faculty efforts with the institutional objectives?

Although the discussion here is presented in the context of academic evaluations, it is equally applicable to other domains where numerical scores are given to a set of criteria. Examples may be found among rankings of airlines, sports clubs and their players, movies, and consumer products.

2. Scope, purpose, and contribution

Notwithstanding the extensive literature on quantitative faculty evaluations, our work focuses on a rather unique aspect of the practice. We would like to investigate the effects of selecting the weights on the individual scores and rankings. Our approach is based on a formalism given in the following section. We then make use of Linear Programming (LP) optimization techniques. The versatility of our approach is demonstrated by several extensions. Our interest is not on single faculty members, but rather on the systems aspects of the practice. That is, how do the weights affect the set of professors, their relative rankings, and their prospects to be the best or words performer, and how sensitive is the evaluation system to the selection of the weights? We also ask how much leverage administration has by manipulating the assignment of the weights. If the rankings are sensitive to the selection of the weights, then administration wields great power and influence over the outcomes. This is especially important if these rankings are used for personnel decisions such as promotions or firings. We would like to identify the circumstances under which the administration has greater influence over the rankings.

It should be noted that scores need not be higher-the-better type. All arguments are equally applicable to the lower-the-better type as well. A simple way to implement this is to consider the scores as penalties. Than any numerical mechanism that identifies the best is equally applicable to identifying the worst. In plain terms,

quantitative evaluation schemes that are used for promotion are symmetrically and equally applicable to schemes employed for firing.

2.1. The formalism

In this work, the approach used to investigate the effects of weight assignments is based on formal (numerical or quantitative) modeling considerations. The formalism is presented in terms of professors and rankings of three criteria: teaching, research, and service. Of course, the formalism may be applied to other domains, such as product rankings based on customer reviews, and may include fewer, or more likely, more criteria.

Table 1. Nomenclature

Parameters		
N	Number of professors	
M	Number of criteria (here we use $M = 3$ i.e., teaching, research, service)	
B	*Big M*, a sufficiently large number	
Sets		
$P = \{P_i\}$	Professors	$i \in \{1,2,\dots,N\}$
$C = \{C_j\}$	Criteria	$j \in \{1,2,\dots,M\}$
Variables		
s_{ij}	Score of professor i criteria j	$\forall i, \forall j$
v_i	Vector of scores s_{ij} of professor i	$\forall i$
S_i	Aggregate score of professor i	$\forall i$
X_i	Binary variable: 1 if Professor i has the highest aggregate score, 0 otherwise	$\forall i$
Y_i	Binary variable: 1 if Professor i has the lowest aggregate score, 0 otherwise	$\forall i$
I_n	Index of professor who is n-th best in ranking	for $n = 1,2,\dots,N$
Performance indicators		
L	The lowest aggregate score of all professors	
H	The highest aggregate score of all professors	
δ_1	Fraction (density) of professors who could achieve the top aggregate score by manipulating the weights	
δ_k	Fraction (density) of professors who could achieve the top k aggregate scores by manipulating the weights[a]	
Decision variables		
w_i	Weight assigned to criteria j	$\forall j$

[a]See Section 6 for a detailed description.

Given the raw scores of professors for each of the criteria, we are interested in the relative ranking of the aggregate scores $\boldsymbol{S_i}$ of the set of professors based on the different assignment of weights. That is, how are the aggregate scores $\boldsymbol{S_i}$ affected by the selection of the weights? Is it possible to manipulate the weights so that, say, a given professor has the highest score? Moreover, with $\boldsymbol{N}$ professors, there are $\boldsymbol{N}!$ possible rankings, or orderings. Is it possible to achieve a given ordering by carefully choosing the weights appropriately? Regarding these inquiries, let us define a system parameter, $\boldsymbol{\delta_1}$.We name this parameter $\boldsymbol{\delta_1}$ since it is a type of density. Specifically, $\boldsymbol{\delta_1}$ is the fraction of the professors that could be placed as the highest aggregate scoring professor. The smallest possible value of $\boldsymbol{\delta_1}$ is $\boldsymbol{1/N}$. This corresponds to the case that one professor dominates all others in every criterion. The largest possible value of $\boldsymbol{\delta_1}$ is 1. This corresponds to the case where any of the $\boldsymbol{N}$ professors may be set to have the highest aggregates score. Clearly, the density $\boldsymbol{\delta_1}$ is determined by the

set of individual raw scores of the professors in each of the criteria, the s_{ij}'s. As such we view the densities as the properties of the system, i.e., of the totality of the scores, rather than the performance of any one of the professors. Equivalently, the density may be interpreted as a probability. It is the probability that a randomly chosen professor has the potential to achieve the highest aggregate score.

3. Numerical experiments

In order to gain deeper insights into the properties of the evaluation system and the presented formalism, we perform several numerical experiments. For a given set of individual scores (s_{ij}), we compute the aggregate scores for all professors $k = 1,2, \dots, N$. We then find the professor with the highest aggregate score. We span the solution space by inspecting all possible combinations of the weights. We count how many of the professors may attain the best score by manipulating the weights in their favor.

3.1. The effect of weights on the determination of the best professor

Let us consider an example with scores as given in the table below (Table 2). We have $N = 5$ professors who are evaluated by $M = 3$ criteria. Each professor has a raw score between 0 and 1 for each criterion.

Table 2. An example: the scores s_{ij} are given for the case $M = 3$ and $N = 5$

Criteria	Professor 1	Professor 2	Professor 3	Professor 4	Professor 5
1	0.65	0.65	0.07	0.65	0.05
2	0.15	0.65	0.80	0.75	0.83
3	0.60	0.25	0.60	0.07	0.57

Given the data, we may pick a set of weights and see which professor attains the highest aggregate score. The set of possible weights constitutes our solution space. The solution space is spanned by w_1, w_2, and w_3. Since $w_3 = 1 - w_1 - w_2$ we may show the entire solution space as a two-dimensional area on the (w_1, w_2) space. Moreover, since $w_1 + w_2$ may not exceed 1, the solution space is a triangular region in the first quadrant of the (w_1, w_2) space. Different points in this space correspond to different weights. Such a space is shown in Fig. 1 below. We paint each point according to which of the five professors attains the highest score.

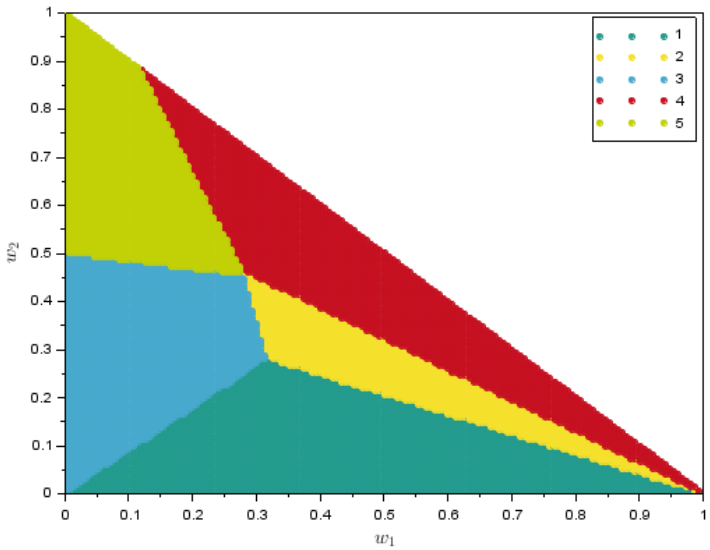

Figure 1. Regions where each professor attains the highest score.

The solution space has five distinct regions, each corresponding to a different professor. Clearly, it is not necessary that every professor attains the highest aggregate score. That is, there may be cases where the

number of regions is strictly less than the number of professors. In the extreme case, suppose one of the professors receives nonnegative scores in each of the criteria, while the remaining $N-1$ professors all receive zero scores. Then, the entire solution space would be dominated by that one professor with nonzero scores.

Fig. 1 provides further insights into the effects of choosing the weights. Interestingly, there is a point, close to (0.3, 0.5) where four separate regions meet. This point corresponds to (approximately) $w_1 = 0.3$, $w_2 = 0.5$, and $w_3 = 0.2$. All but Professor 1 become the highest scoring person in the vicinity of this point. In another sense, the selection of the "best" professor is highly sensitive to the selection of the weights. By a small change, we may allow any one of the four professors to be the best. As mentioned, the same argument holds for the "worst" professor. If evaluations were used for firing the worst performer, by slightly perturbing the weights, we may select any one of the professors to be fired.

We would like to ascertain if a given professor could attain the highest aggregate score without the need to scan the entire solution space. In the next section, we present a formal model that would address this question.

4. A linear programming (lp) formulation

The computation of the weights that leads to prescribed circumstances is easily accomplished by a simple linear programming (LP) implementation. Actually, the problem at hand has sufficient properties that may lend itself to closed-form solutions. However, for the sake of generality, as an initial attempt, we are lured by the expediency of the LP approach.

Let us consider the case where Professor k is to have the highest aggregate score. We consider the following formulation.

$$max\ S_k \tag{1}$$

$subject\ to$:

$$S_i = \sum_{j=1}^{M} w_j\, s_{ij} \qquad \forall i \tag{2}$$

$$S_k \geq S_i \qquad \forall i \tag{3}$$

$$\sum_{j=1}^{M} w_j = 1 \tag{4}$$

$$w_j \geq 0 \qquad \forall j \tag{5}$$

Equation (2) defines aggregate scores as linear combinations of the individual scores, while inequality (3) forces S_k to be the maximum of all scores. Equation (4) is the normalizing condition, while inequality (5) forces the weights to be nonnegative. This model, if a feasible solution exists, gives the set of weights $\{w_1, w_2, \dots, w_M\}$ that maximizes the aggregate score of Professor k (that is, S_k).

Whether or not a given professor may receive the highest aggregate score may be discovered by solving the LP described above. A feasible solution would indicate that it is possible for the given professor to attain the highest aggregate score. While the LP will yield only a single feasible point in the weight space, more work is needed to plot the entire region where that professor receives the highest aggregate score. Theoretical considerations in LP prescribe the so-called *sensitivity analysis* as a means to discover the entire region throughout which the solution remains optimal. The LP view of the phenomena and the accompanying

formulation provide additional benefits. For example, LP guarantees that if a feasible region exists, it is contained in a contiguous linear convex region, referred to as a simplex. Thus, we know that the picture we have from a single example above (see Fig. 1) is not an exception, but a typical case. In three dimensions (three decision variables) the regions will be convex areas bounded by lines. Although an important and worthy endeavor, we will not pursue sensitivity analysis in this paper, but rather dwell on the insights from the formulation and the meaning of evaluation.

5. Performance indicator densities

In previous sections we defined a key performance indicator. The density δ_1 is the fraction of professors who could receive the top score by suitably selecting the weights. The density δ_1 clearly takes values in interval $[1/N, 1]$. We experiment with randomly generated cases to gain further insights into this performance indicator. Numerical experiments were based on generating random data and analyzing the results of the LP solutions. Sets of normally distributed scores (s_{ij}) were generated. Each set contained 250 replications. The density δ_1 was computed for different numbers of professors ($N = 5, 10, 15, 20, 25, 30, 35, 40$) and three criteria ($M = 3$). Fig. 2 below shows the distribution of the density δ_1.

As seen, when there is a large number of professors, there are fewer cases where a given professor may attain the best score. For $N > 30$, δ_1 is about 20%. However, when the number of professors is small, for $N = 5$, then the density is much higher (δ_1 is about 60%). Fig. 2 gives box plots for the density δ_1. Not only the median is shown as a bar, but also the quartiles are indicated [10]. It may be argued that if the number of professors is large, there are always a few "star" professors who excel at many criteria, and thus dominate most other professors in ranking. This property seems plausible. Consider the case where there are much more criteria than professors. It stands to reason then that it would be easier to pick different criteria that make a given professor receive the highest aggregate score.

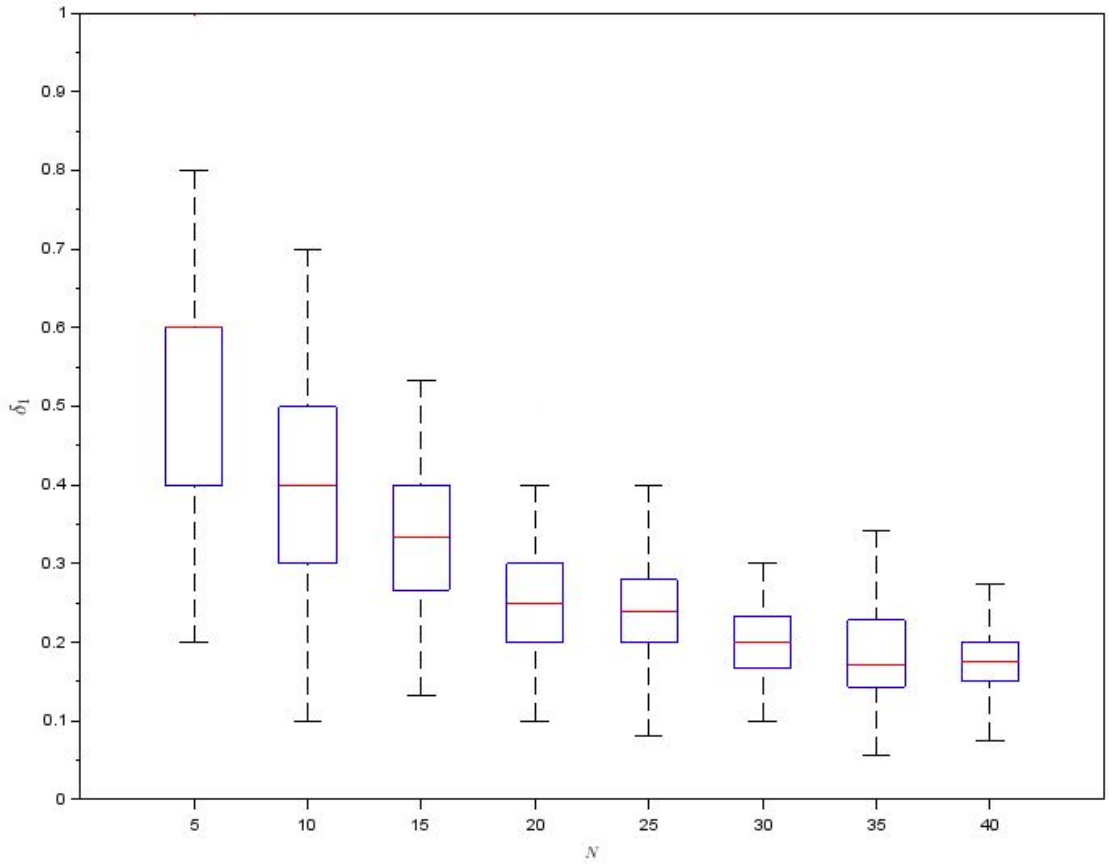

Figure 2. The effect of the number of professors on the density δ_1 (M= 3, 250 replications of each experiment)

It is not unreasonable that not all professors may reach the top rank. After all, there will be cases where $v_k \geq v_l$, that is when all of the scores of professor k are at least as large as that of professor l, that is, when $s_{kj} \geq s_{lj}$. Then, irrespective of the weights, professor P_k will have a higher aggregate score compared to professor P_l. In the language of mathematical programming, we say that P_l is dominated (by P_k). The fewer the criteria, the higher the chance of dominated elements. After all, if we had only one criterion, then only one professor could achieve the top rank, and all others would be dominated. Similarly, a large pool of professors would be expected to contain many dominated ones.

Here, we would like to point out that our assumption that scores (s_{ij}) are normally distributed is only an introductory consideration. Not having any prior insights, we simply assume a normal and statistically independent distribution for each raw score. It is entirely plausible that over time, a given evaluation process would lead the professors to adjust their efforts in preferable directions. In all likelihood, there would be professors who naturally gravitate towards one or more of the criteria with higher weights. These considerations will be revisited later. For the time being, our objective is to gain preliminary insights into the systems aspects of such quantitative weighted evaluation practices. Preliminary work shows that the insights listed above are relatively independent of the exact distribution of the raw scores as long as the scores are statistically independent. This claim seems plausible, since we are interested in orderings and not the exact amount of differences between scores.

We next investigate the effect of the number of criteria on the distribution of the density δ_1. A similar set of random data is generated following the normal distribution assumptions. This time, number of criteria (M) is changed while the number of professors (N) is fixed at 20. As the number of criteria and the number of corresponding weights (decision variables) increase, system becomes more malleable. As expected, the possibility of any professor achieving the best score grows rapidly. Fig. 3 summarizes the numerical experiment results. With eight different criteria (M = 8), it seems about 80% of the professors may achieve the highest score by so manipulating the weights.

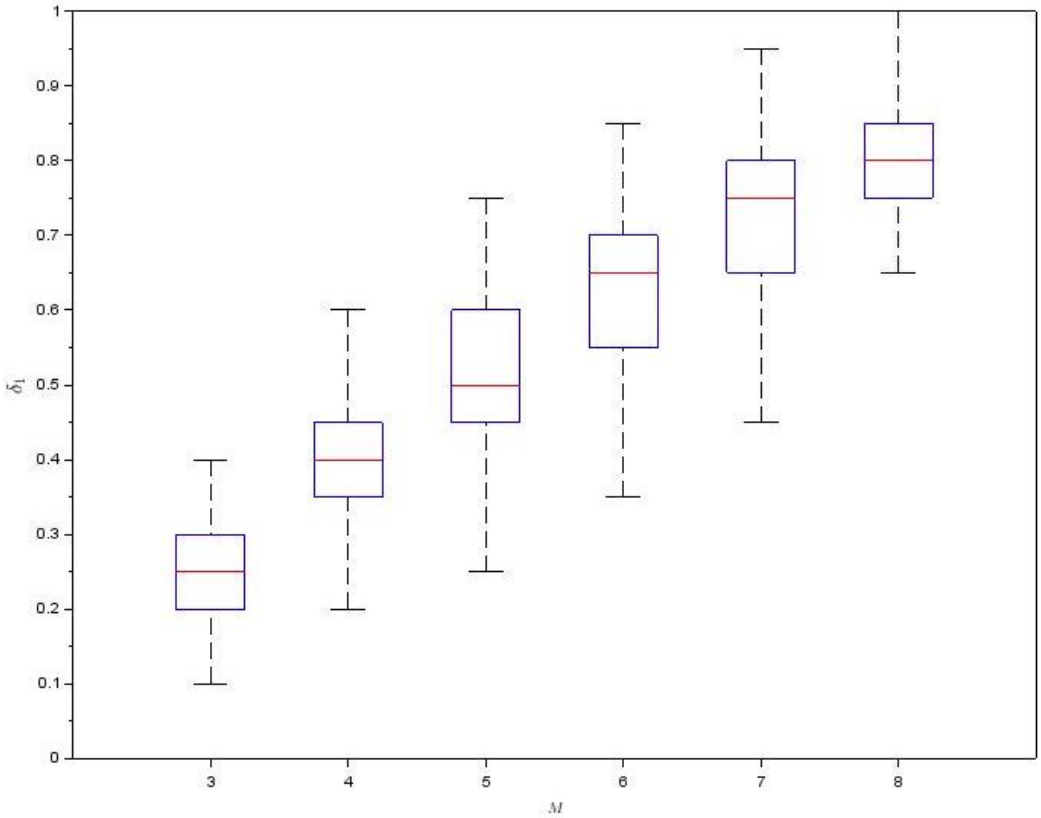

Figure 3. The effect of the number of criteria on the density δ_1 (N=20, 250 replications of each experiment)

It is somewhat surprising, however, that the density δ_1 exceeds 80% with only eight criteria. That is, with eight criteria, it is possible to practically place any professor from a group of 20 as the top performer. This gives the decision maker who assigns the weigh a considerable amount of power over the rankings. Once again, we note that the analysis works symmetrically for ranking the worst performers. As such, the insights are equally valid when we consider firing decisions versus hiring or promotion decisions.

6. The effect of weights on the order of aggregate scores

The formal model given by equations (2) - (5) seeks to bring only one professor to the top rank. We may also be interested to see if a given ordering of the professors is possible. That is, not only do we want a certain professor to have the highest score, but we also want to place certain professors in second place, third place, etc. This requires a few changes in the initial LP. In the most general case, we would like to see if the top k rankings could be assigned to k specific professors. We define the density δ_k as the fraction of $\binom{N}{k}$ orderings of the best k professors that can be achieved by manipulating the weights.

Let us call the set of all k rankings $\underline{\Omega_k}$, with elements $< I_1, I_2, \dots, I_k >$ where I_n is the index of the professor whose score is to be the n-th best among all professors. The set $\underline{\Omega_k}$ has $\binom{N}{k}$ elements. The set $\underline{\Omega_N}$ has cardinality $N!$. We use a second subscript to differentiate among the $\binom{N}{k}$ orderings, that is, we let $\Omega_{k,l}$ denote the l-th such ordering, where $l = 1,2, \dots, \binom{N}{k}$.

We no longer want a specific professor to have the highest score, so we remove inequality (3) and replace it with the following set of inequalities.

$$S_{I_n} \geq S_{I_{n+1}} \qquad \text{for } n = 1,2, \dots, k-1 \tag{6}$$

Since we are only interested in whether or not such a set of weights exists that yields this particular ordering, the objective function is inconsequential. Some software packages nonetheless require a dummy objective function, such as the one below.

$$max \ \ 1 \tag{7}$$

An alternative objective may be to maximize the score of the top-ranking professor.

$$max \ \ S_{I_1} \tag{8}$$

With these changes, we may use the LP formulation to see if the weights may be manipulated to achieve the given ordering of the professors. A feasible solution means that there exists weight that would yield the desired ordering. As another numerical experiment to gain insights, let us use the following raw scores and compute δ_N, that is δ_5,

Table 3. An example: the scores s_{ij} are given for the case $M = 3$ and $N = 5$

Criteria	Professor 1	Professor 2	Professor 3	Professor 4	Professor 5
1	0.63	0.61	0.55	0.47	0.31
2	0.60	0.38	0.03	0.90	0.68
3	0.10	0.23	0.52	0.31	0.81

The five professors may be ordered in a total of 5! = 120 different ways. Again, we do not expect all such orderings to be feasible. The experiment yields 26 regions as depicted in Fig. 4 where each color represents different ordering. In this example, a little over 20%, (i.e., $\delta_5 = \frac{26}{120}$) of the orderings turn out to be feasible. Note that here δ_1 = 1 since any of the professors may achieve the top score.

Once again, as seen in Fig. 4, we observe a point near (0.6, 0.1) where many regions meet. The ordering of the professors becomes exceedingly sensitive to the selections of the weights near this point, (i.e., near $w_1 = 0.6$, $w_2 = 0.1$, $w_3 = 0.3$). A close examination reveals that there are actually several small regions clustered around this point, besides the several wedge-shaped regions.

The rank orderings are important even if the best performers remain unchanged. Many institutions have uncontested *super*-performers, who routinely receive rewards. Then the competition is focused on the second tier of faculty to jostle for position among this group. While not super-performers, they nonetheless compete for rewards such as raises, sabbatical leaves, equipment and supplies, graduate students, etc.

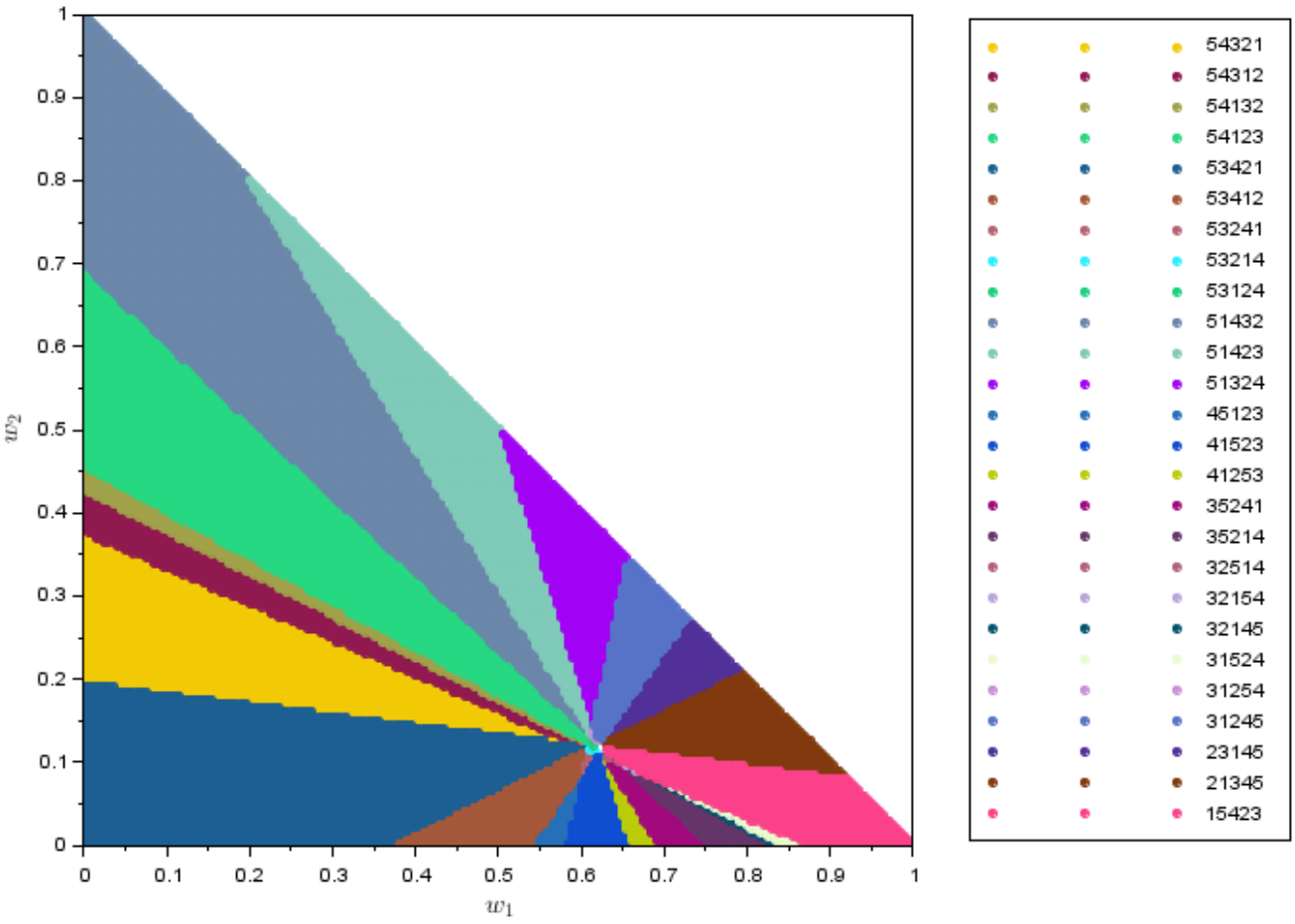

Figure 4. Feasible regions for different rankings (Data from Table 3)

In the example discussed, we see that δ_N is a little over 20%. The two density measures δ_N and δ_1 are related. If δ_N is high, one would expect δ_1 to also be high, since the more likely to achieve any given ordering implies that it would be easier to place any given professor to be the best.

7. The effect of weights on the range of aggregate scores

The range of aggregate scores is a good indicator of the spread of the performance of the professors. We next ask how much may the range of scores be affected by the choice of the weights. That is, do the selection of the

weights have a significant impact on the difference between the best and worst aggregate scores?

Once again, we take advantage of the versatility of the LP formulation. We make use of binary variables X_i and Y_i which are defined for each professor i. X_i is set to 1 if Professor i has the highest aggregate score, and 0 otherwise. Similarly, Y_i is set to 1 if Professor i has the lowest aggregates score, and 0 otherwise. The decision variables X_i and Y_i may be regarded as indicators.

The constraints to tease out the highest and lowest scores are given below. The formulation extends the basic formulation given Section 2.1, Equations (2, 4, 5) are directly taken from the basic formulation.

$$min/max\ H - L \tag{9}$$

$subject\ to$:

$$S_i = \sum_{j=1}^{M} w_j\, s_{ij} \qquad \forall i \tag{10}$$

$$H \geq S_i \qquad \forall i \tag{11}$$

$$L \leq S_i \qquad \forall i \tag{12}$$

$$H - B(1 - X_i) \leq S_i \qquad \forall i \tag{13}$$

$$L + B(1 - Y_i) \geq S_i \qquad \forall i \tag{14}$$

$$\sum_{i=1}^{N} X_i = 1 \tag{15}$$

$$\sum_{i=1}^{N} Y_i = 1 \tag{16}$$

$$\sum_{j=1}^{M} w_j = 1 \tag{17}$$

$$w_j \geq 0 \qquad \forall j \tag{18}$$

The highest score H must be greater than or equal to all aggregate scores. Equation (11) states this condition. Similarly, the lowest score L must be less than or equal to all aggregate scores, as required by equation (12). Equation (13) forces the variable H to be one of the aggregate scores. Here we make use of the so-called *Big M* method. The parameter B is a sufficiently large number. If Professor i does not have the highest score, then the corresponding binary variable X_i will be zero, and hence, inequality (13) will be satisfied regardless. Here the parameter B need only be larger than 1. If Professor i does have the highest score, then inequality (13) together with inequality (11) will force H to be the score S_i. A similar formulation is used for the lowest score. Inequality (14) along with inequality (12) set L to the lowest aggregate score with the help of the binary decision variables Y_i. Finally, equation (15) allows only one of the binary variables X_i to be 1 and all others zero. Similarly, equation (16) allows only one of the binary variables Y_i to be 1 and all others zero. That is, we pick only one aggregate score to be the highest and only one to be the lowest.

The additional constrains given above allow us to seek the weights that yield the widest and narrowest range of aggregate scores. We use the objective functions to find the widest and narrowest ranges. Once again, we

use the example given by Table 2 to discover that $w_1 = 0.34$, $w_2 = 0.31$, and $w_3 = 0.35$ gives a range of only 0.03. The lowest aggregate score in this case is $S_1 = 0.48$ and the highest aggregate score is $S_2 = 0.51$. The range is practically nonexistent and all professors have almost the same aggregate score. On the other extreme, the weights $w_1 = 0$, $w_2 = 1$, and $w_3 = 0$ give the largest possible range of 0.68, where $S_5 = 0.83$ and $S_1 = 0.15$. Granted, the extreme case of one of the weights equal to 1 and the others zero is rather unrealistic. However, it is nonetheless rather surprising that the range of aggregate scores could be affected so much by the choice of the weights.

A nefarious administrator may be interested not in the order of the entire set of professors, but only in elevating one specific professor from another. Although such intent is to be condemned, we nonetheless would like to know if potential for such activity exists. The versatility of the LP formulation is once again acknowledged. We single out two professors and want to make one have an aggregate score above the other as much as possible. The following objective function maximizes the difference among the aggregate scores of professors P_k and P_l. Of course, scenarios involving more professors are also implemented quite easily by the LP formulation, once again illustrating the versatility and universality of the formulation. We introduce the inequality,

$$S_k > S_l \tag{19}$$

and set the objective function to force the difference to be as large as possible.

$$\max S_k - S_l \tag{20}$$

If a feasible solution exists, the difference of the aggregate scores of the two professors will be as wide as possible.

8. Conclusion

It is quite common that competing entities (professors, products, airlines, etc.) are evaluated based on an aggregate score computed as a weighted sum of individual scores in each of the criteria. We present our work in the context of professors, along with the common criteria of *teaching*, *research*, and *service*. We question the leverage a decision maker who determines the weights would have on the outcome of the rankings. Our approach is based on numerical examples and formal linear programming (LP) considerations.

We find that the nature of the phenomenon of such evaluation is indeed sensitive to the selection of weights. We show that in many cases, small perturbations to the weights may result in many different rankings of the faculty. Albeit being perceived as embodying numerical precision, the practice may lead to unintentional and rather unsubstantiated erroneous conclusions. Moreover, the practice may be vulnerable to intentional manipulation by those who set the weights.

Acknowledging the findings presented here would lead to more fair practices. As such, this work suggests several additions to the common practices. The process may benefit from the reporting of anonymized professor raw scores and the solution space. Cases where small perturbations have large effects may be discounted. If such cases persist, the institution may wish to alter its evaluation strategy from a strict numerical ranking towards a more constraints-based strategy. That is, the institution may choose to set minimum achievement levels, and regard any professor who meets or exceeds these thresholds to be satisfactory.

This work encourages further investigations at a wider scope. The systems view of the weights and the grading practice encapsulates information regarding the distribution of talent and capabilities among the faculty members. Issues such as how similar or dissimilar the faculty members are relating back to the level of

diversity versus uniformity of the set. Along the same lines, data from successive years contain information regarding whether the population characteristics converge or diverge. Divergence in this sense expands the collective range and domain of the talent, skill, and expertise of the institution. Convergence implies conformity and regression towards a singularity, towards a setting where all members are rather similar in their skills and talents. Which of these trends is preferable and encouraged depends on the institution and the specific nature of the evaluation? If done intentionally and calculatingly, these are the tools to guide the body of faculty. However, there is also a chance to adopt evaluation policies that seemingly implement the traditionally popular mechanisms, but result in rather unintentional consequences. Knowing the expected effects of the evaluation mechanism is thus in the interest of the institution. This study provides a way to measure diversity as the possible range of scores, as given in Section 7. That is, rather than measuring the range of achieved scores, the institution may benefit from comparing the range of the widest and narrowest possible range.

The work also points to promising future directions. A mathematical study to further scrutinize the properties of densities δ_k would be beneficial, especially if it yields closed-form solutions. And finally, the versatility of the LP approach developed in this study may lead to the investigation of related performance measures. In this sense, the study is regarded as having presented a potential general methodology for future investigations. Finally, the concerns addressed here are applicable to other domains of evaluations and rankings, such as the rankings of airlines, sports clubs, movies, and consumer products. Any extensions of the results here to these fields are encouraged.

References

[1] A. Martin, T. Miranda Lakshmi, and V. Prasanna Venkatesan, "A Study on Evaluation Metrics for Multi Criteria Decision Making (MCDM) Methods - TOPSIS, COPRAS & GRA", *International Journal of Computing Algorithm*, vol. 7, no. 1, pp. 29-37, 2018.

[2] B. Gunn, "Salary administration in the management systems of higher education", *Innovative Higher Education*, vol. 13, no. 2, pp. 117-146, 1989.

[3] B. Younes, Faculty evaluation: Towards a happy balance between competing values. World Transactions on Engineering and Technology Education, vol. 2, no. 1, pp. 117-120, 2003.

[4] C. A. B. Bana e Costa and M. D. Oliveira, "A multicriteria decision analysis model for faculty evaluation", *Omega*, vol. 40, no. 4, pp. 424-436, 2012.

[5] C. L. Colbeck, "Integration: Evaluating faculty work as a whole", *New Directions for Institutional Research*, vol. 2002, no. 114, pp. 43-52, 2002.

[6] E. Thanassoulis, P. Dey, K. Petridis, I. Goniadis and A. Georgiou, "Evaluating higher education teaching performance using combined analytic hierarchy process and data envelopment analysis", *Journal of the Operational Research Society*, vol. 68, no. 4, pp. 431-445, 2017.

[7] I. F. Jaramillo, R. Pico and C. De La Plata, "A Model for Faculty Evaluation in Higher Education Ecuadorian through Multi-Criteria Decision Analysis", *Indian Journal of Science and Technology*, vol. 10, no. 18, pp. 1-8, 2017.

[8] J. Centra, *How Universities Evaluate Faculty Performance: A Survey of Department Heads*. Graduate Record Examination Program, 1977.

[9] J. P. Runtuwene, I. R. Tangkawarow, and M. T. Parinsi, "Analytic Hierarchy Process (AHP) Methods For Evaluation of Teacher Quality". In International Conference on Science and Technology (ICST 2018). Atlantis Press. 2018.

[10] J.W., Tukey, Exploratory data analysis, vol. 2. 1977.

[11] K. Redmon, "ERIC Review Faculty Evaluation: A Response to Competing Values", *Community College Review*, vol. 27, no. 1, pp. 57-71, 1999.

[12] L. Romney and C. Manning, *Faculty activity analysis*. Boulder, Colo.: National Center for Higher Education Management Systems at Western Interstate Commission for Higher Education, 1971.

[13] L. S. Root, “Faculty evaluation: Reliability of peer assessments of research, teaching, and service”. Research in Higher Education, vol. 26, no. 1, pp. 71-84, 1987.

[14] M. bin Othman and S. bin Abdullah, "AHP based academic performance scoresheet (APS) for holistic assessment of academician achievements", in *Proceedings of the International Symposium on the Analytic Hierarchy Process*, 2013.

[15] M. Collan, J. Stoklasa and J. Talasova, "On academic faculty evaluation systems - more than just simple benchmarking", *International Journal of Process Management and Benchmarking*, vol. 4, no. 4, p. 437, 2014.

[16] R. Arreola, Developing a comprehensive faculty evaluation system. Bolton, MA: Anker Pub. Co., 2000.

[17] R. Bauwens, M. Audenaert, J. Huisman and A. Decramer, "Performance management fairness and burnout: implications for organizational citizenship behaviors", *Studies in Higher Education*, vol. 44, no. 3, pp. 584-598, 2017.

[18] R. Ebel, Essentials of educational measurement. Englewood Cliffs, N.J.: Prentice-Hall, 1972.

[19] S. Cadez, V. Dimovski and M. Zaman Groff, "Research, teaching and performance evaluation in academia: the salience of quality", *Studies in Higher Education*, vol. 42, no. 8, pp. 1455-1473, 2015.