ORIGINAL RESEARCH

Pooling Rankings to Obtain a Set of Scores for a Composite Indicator of Erasmus + Mobility Effects

Luigi Fabbris[1] · **Manuela Scioni**[1]

Abstract

In this paper, we study how to assign weights to a set of evaluations obtained at the end of an international mobility experience in order to aggregate them into a composite indicator. The mobility experience was evaluated by three categories of actor: the participant; the school or company sending the participant; and the school or company hosting the participant. We estimated the weights starting from the assessors' mutual evaluations of the beneficiaries of the mobility experiences. In particular, the aim of the paper was to compare two strategies for estimating the weights: (1) a weighted function of the univariate rank distribution of frequencies; and (2) the normalised elements of the first eigenvector of the dominance matrix computed by mediating the actors' dominance matrices derived from the rankings of mobility beneficiaries. Variants of the two strategies were also introduced. Even though each strategy had different assumptions, the analyses produced several important findings. First, the optimum weighting model depends on the loss function used to evaluate the quality of the results. In particular a *between-ranking variability function* favours both univariate and unweighted multivariate models, while a *bias-based function* favours weighted multivariate models. Second, in both univariate and multivariate analyses, the application of *rank-order-centroid* and *rank-reciprocal rules* give more accurate results than both linear and exponential rules.

Keywords Aggregation of rankings · Score estimation · Composite indicator · Weighting system · Dominance matrix

1 Introduction

ROI-MOB (Measuring return on investment from EU VET mobility) is a European project that aims to represent the final outcomes of an Erasmus + mobility experience using a single, complex indicator (Fabbris and Boetti 2019). ROI-MOB focuses on the Vocational

✉ Manuela Scioni
  manuela.scioni@unipd.it

  Luigi Fabbris
  luigi.fabbris@unipd.it

1 Department of Statistical Sciences, University of Padova, Via Cesare Battisti 241, 35121 Padua, Italy

Education and Training (VET) experience, which consists of an internship undertaken by a student or apprentice at a workplace in another country. Schools, companies, or other intermediate bodies organize, in agreement with the hosting bodies, the internship of individuals or small groups of participants. The indicator measures the quality of this complex phenomenon jointly experienced by the young participant and all other bodies involved in the process. The indicator is a composite index of elementary indicators, each capturing the viewpoint of an assessor, aggregated using an appropriate 'importance' weighting. In particular, the surveys collected data from four groups of assessors: (i) the participants directly involved in VET international mobility, either residing or hosted in one of four countries (Germany, Italy, Portugal, and Spain); (ii) the schools and training centres facilitating the participation of students or apprentices, either by sending them abroad or hosting them; (iii) the companies and public bodies that sent or hosted the participants; and (iv) representatives of the national or European institutions promoting VET international mobility. For the sake of simplicity, only the data from the first three groups of assessors are processed in this paper.

This paper aims to estimate the importance weights to attach to the final evaluations from three groups of assessors surveyed after a set of mobility experiences that occurred in 2017 and 2018 in the four focal countries. For each assessor, the final evaluation is assumed to be an overall evaluation measure of the outcomes of the pertinent VET mobility experience. To estimate the weights, we applied a set of procedures to the rankings of beneficiaries from the VET international mobility scheme, as expressed by the assessors, and defined an optimum way to transform rankings into scores. The procedures rely on the hypothesis that the mobility experience is more important for those categories of assessors who receive greater benefits and, therefore, the evaluation of the experience of these categories must weigh more in the composite indicator representing the overall evaluation.

Technically, the ROI-MOB indicator is a combination of the final evaluations independently given by the mobility actors, weighted according to the recognised 'benefit' the actor received from the mobility experience. All actors are assumed to be equally informed of the process. This prevents the argument be made that the weights attached to the experience assessments should be proportional to the level of information possessed by the assessor, which is unknown. However, we do know the level of benefit the assessors obtain through the process.

In the following, we deal just with the aim of estimating the weights for the indicator construction, which we will call *scores* so as not to confuse them with the weights that will be used in their estimation. The rest of the paper is organised as follows. Section 2 describes in detail the available data, the analytical models used in the estimation of weights, and the criteria for model selection. Section 3 describes the results, and these are then discussed in Sect. 4.

2 Data and Methods

2.1 The Data

The data came from three surveys conducted between March and August 2018 in four European countries (Germany, Italy Portugal, and Spain). The questionnaires were sent to samples of participants, schools, and companies whose email addresses were provided by the project partners. All participants, companies, and schools involved in a VET mobility

experience organised by one of the project partners in the last two years were included in the sample. Each sample aimed to represent a national context. After two pilot rounds of data collection, the questionnaires, written in the four national languages of the project (Italian, German, Portuguese, and Spanish), plus English, were administrated to over 5,000 potential respondents through a Computer Assisted Web-based Interviewing (CAWI) system. The participation-level was good, with a response rate of 31% (1,545 questionnaires).

The national representativeness of the samples does not provide European representativeness in the overall sample. However, the presence in the sample of both Mediterranean and Central Europe countries allows for broad cross-national inference. Moreover, the four focal countries account for a substantial proportion of the international students and trainees mobilised by the Erasmus + Programme (Fabbris and Boetti 2019).

The basic questions posed in the surveys to evaluate the ranking of beneficiaries slightly differed according to the mobility actor. The questions posed only to participants was:

- *'Finally, which are the two categories that get the highest benefits from Erasmus + mobility? (Please, click the first and the second category of possible recipients)'*. The option was given to specify a first and second selection out of five possible beneficiaries: students/apprentices, schools and training centres, companies (both sending and hosting), the labour market, and the European Union as an institution. The beneficiaries were ranked on a three-level ordinal scale.

    The question posed only to the schools and companies was:
- *'Finally, which are the categories of possible recipients that get the highest benefits and the ones that get the lowest ones from Erasmus + mobility? Please, order the categories from 1 (highest) to 5 (lowest benefits)'*. The same categories presented to the participants were used here. The beneficiaries were ranked on a five-level ordinal scale.

The collected data contained nonresponses. In some cases, people did not respond to the question at all; in other cases, people indicated just the first position or just few initial positions. All given rankings, even those that were incomplete, were processed.

It is worth highlighting that the three categories of assessors evaluated themselves as a possible recipient of mobility benefits. Thus, when a respondent evaluated the relative positions of the mobility actors, they were expected to operate a sequence of mental comparisons, first pinpointing the top beneficiary and then selecting a second best among the residual categories, then a third, and so on. All the possible beneficiaries were well known to the assessors as actors of the represented mobility process.

2.2 Models and Methods

In the following, we deal with the evaluation systems in which each unit of a sample of n independent assessors ranks A' alternatives from a set of A total alternatives, $A' \leq A$, from the first to last possible place. Our aim is to find an optimum strategy to aggregate the n rankings and transform them into scores to assign to the A' alternatives.[1] Without a loss of generality, we assume that—as occurred in our surveys—all alternatives are ranked, e.g. $A' = A$.

---

[1] For a survey of the literature on the transformation of ordinal into interval data the reader is referred, among the others, to Manisera and Vezzoli (2012), Ciavolino and Carpita (2015) and Zanarotti and Pagani (2015).

Let us consider the information contained in the rankings given by a group of assessors. To estimate the scores, the information can be conceived at different levels:

- We can use just the top position of each distribution. Thus, the score estimates are proportional to the frequencies of the top positions obtained by the A alternatives. This estimation method uses basic information and does not require particular statistical techniques. In the following, this procedure is called 'first position'.
- We can deduce scores from the averages of frequency distributions. This estimation method uses the full frequency distribution as if each ranking was an empirical occurrence of a random variable distributed with mean $(A + 1)/2$ and variance $(A^2 - 1)/12$. Henceforth, we will call this method the 'univariate' procedure.
- We can model the relations between the $A(A - 1)/2$ distinct pairs of alternatives by comparing their ranking positions and defining the level of 'dominance' of alternative a over alternative b. Thus, this weighting procedure is called 'bivariate'. This estimation method requires the construction of a 'dominance matrix', which reflects the bivariate relationships between alternatives. The way such a matrix is defined and how it can be processed for weight estimation purposes is described later in this section.

The univariate approach uses the (absolute or relative) frequency distribution of the rankings given by the n respondents. For the a th alternative, the absolute frequency distribution is the vector $\mathbf{f}_a = \{f_{a1}, ..., f_{ah}, ..., f_{aA}\}$ whose elements add up to n. To give an example, f_{a1} is the number of respondents who put the a th alternative in the first position. The score for alternative a, Y_a, is obtained as a linear combination of the frequencies at all rank positions of that alternative,

$$Y_a = \sum_{i=1}^{A} f_{ai} W_i \qquad \left(W_i \geq 0; \ a = 1, \dots, A\right) \qquad (1)$$

where: f_{ai} is the frequency of assessors who placed alternative a in position $i(i = 1, ..., A)$; and W_i is a weight assigned to rank i, such that $\sum_{i=1}^{A} W_i = 1$. For the sake of simplicity, the score is assumed to be normalised, so as to add up to one across all alternatives, $\sum_{a=1}^{A} Y_a = 1$ and $0 \leq Y_a \leq 1$.

For score estimation purposes, the bivariate approach refers to the analysis of dominances, tournaments, or round-robins (Ahn and Park 2008). This analysis requires a three-step procedure:

1) Estimation of a dominance matrix, $\mathbf{P} = \{p_{ab}(a \neq b = 1, ..., A); p_{aa} = 0 \ \forall a, b\}$, through the rankings expressed by the respondents. The dominance relations can be recorded in a 'generalised' tournament matrix, whose entries vary between zero and one, with extremes included (Moon and Pullman 1970; Tanino 1988). $\mathbf{P}$ is a square skew-symmetric matrix of cardinality A with $p_{ab} = 1 - p_{ba}(a \neq b = 1, ..., A)$ and $p_{aa} = 0$ for $(a = 1, ..., A)$. The entry p_{ab} can be interpreted as a measure of the dominance of a over b, and the more this measure is close to one, its maximum, the larger the score assigned to a with respect to b.

2) The computation of the right eigenvector w corresponding to the first eigenvalue γ of matrix $\mathbf{P}$ is as follows:

$$\mathbf{P}\mathbf{w} = \gamma \mathbf{w} \tag{2}$$

imposing $w'w = 1$. The a th entry of this eigenvector is proportional to the score of the a th alternative.

3) Normalisation of the values of $\mathbf{w} = \{w_a (a = 1, ..., A)\}$ is accomplished by dividing each element with the sum of the A elements of the eigenvector $\sum_a^A w_a$ so that the normalised values w_a^* add up to one across the A alternatives, $(\sum_a^A w_a^* = 1)$. So, in our case, the normalised values estimate how much the actors benefitted from the international mobility on a 0–1 scale so that the total benefit of the $A = 5$ possible beneficiaries is 100%.

Therefore, for bivariate analysis, we estimate the dominance relations through the following formula:

$$P_{ab}^* = \sum_{i=1}^{A-1} \frac{n(a > b | a = i) W_i}{m} = \sum_{i=1}^{A-1} f(a > b | a = i) W_i \qquad (a \neq b = 1, \dots, A) \tag{3}$$

where $n(a > b | a = i)$ denotes the number of times the alternative a is at rank i and the alternative b follows in the sequence, say from $(i + 1)$ to A; m is the total number of comparisons between a and b; and W_i is a weight associated with the number of times a dominates b when alternative a is at rank $i (i = 1, ..., A - 1)$.

For P_{ab}^* to be an element of the $\mathbf{P}$ matrix it requires further refinement:

$$p_{ab} = P_{ab}^* / (P_{ab}^* + P_{ba}^*) \qquad (a \neq b = 1, \dots, A)$$

so that $p_{ab} = 1 - p_{ba}$.

In defining the weights, both in the univariate and the bivariate approach, we will refer to various scoring functions. Let us start with the plurality and the Borda rules (Borda 1784; Brams and Fishburn 2002). The plurality rule is the simplest way of weighting since it assumes that $W_1 = 1$ and $W_h = 0$ for $h = 2, ..., A$. This rule is equivalent to considering only the first position (i.e. the univariate approach) and only the comparisons involving the first ranking position for each alternative in the bivariate approach. In these cases, the scores assigned to the A alternatives are given by the vector $\mathbf{y} = \{\hat{Y}_1, ..., \hat{Y}_a, ..., \hat{Y}_A\}$ in which the score estimate of alternative a is given simply by the relative frequency of the first positions assigned to the alternative a by the n assessors.

The application of the Borda rule implies different weights, $W_i (i = 1, ..., A)$, according to the position of the alternative a in the ranking. For a univariate analysis, the rule assigns a weight of A to the first position, $A - 1$ to the second position, and so on until the weight of one is given to the last position. So, the standardised form of weight W_i is a linear function of the distance of alternative a from the first position of the ranking:

$$W_i' = W'(a = i) = \frac{A - i + 1}{\sum_j^A A - j + 1} = \frac{2(A - i + 1)}{A(A + 1)} \qquad (i = 1, \dots, A) \tag{4}$$

which is decreasing in i, the position of the alternative in the ranking. The weight has a maximum of $2/(A + 1)$, corresponding to the first position, and a minimum of $2/(A^2 + A)$, corresponding to the bottom position:

$$\frac{2}{A(A+1)} \leq W_i' \leq \frac{2}{A+1}.$$

The weights obtained according to formula (4) are based on the idea that the rank order should be reflected directly in the weights and are known in the literature as rank sum (RS) weights ($W_i' = W_i^{RS}$: Stillwell et al. 1981). Another way for converting the rankings into scores is the rank reciprocal (RR) rule, in which weights are proportional to the reciprocals of the rank order of alternative *I* (Barron 1992; Stillwell et al. 1981):

$$W_i^{RR} = W^{RR}(a = i) = \frac{1/i}{\sum_{j=1}^{A} 1/j} \qquad (i = 1, \dots, A) \tag{5}$$

Variants of the rank reciprocal rule are the so called rank order centroid (ROC), in which weights are proportional to the sum of rank reciprocals from position *i* till the bottom position of the ranking:

$$W_i^{ROC} = W^{ROC}(a = i) = \frac{1}{A} \sum_{i=i}^{A} \frac{1}{i} \qquad (i = 1, \dots, A) \tag{6}$$

and the Sum-Reciprocal (SR) rule as proposed in Danielson and Ekenberg (2014), in which weights are intermediate between the linear and the reciprocal ones:

$$W_i^{SR} = W^{SR}(a = i) = \frac{\frac{1}{i} + \frac{A+1-i}{A}}{\sum_{j=1}^{A} \frac{1}{j} + \left(\frac{A+1-j}{A}\right)} \qquad (i = 1, \dots, A) \tag{7}$$

Both RS and ROC approaches have sound statistical bases, both relying on distributional properties and being subject to random error (Jia et al. 1998). When numerical weights are estimated through functions, instead of directly observed, they are also called surrogate weights (Danielson and Ekenberg 2017).

The literature on how independent rankings can be merged into one or more summarizing objects of the same ordinal (ranking)[2] or metric (score) nature makes reference to the *feasible consensus*, *compromise theory in social choices*, or *visual representations of the rankings in a one-dimensional normalised scale* (Honda et al. 1981; Jensen 1986; Hudry and Monjardet 2010). A metric solution requires the introduction of the notion of distance: the consensus scores are those minimizing the remoteness from the basic rankings, and the remoteness depends on the imposed metric. In what follows, to account for the possible dependence between the solutions and the adopted metrics, we applied various metric rules.

---

[2] In case the consensus estimate is the median ranking—which is *maximum likelihood* (Young 1995)—, it possesses mathematical properties, e.g. the Pareto rule property (whenever all assessors rank one alternative over another, so does the final ('consensus') ranking); the independence of irrelevant alternatives property; as well as the so-called reinforcement rule (if two distinct groups of assessors each reach the same consensus ordering under independent assessment, this ordering is the consensus for the merging of the two groups). It can be concluded that, even if other metrics are plausible (a profile can have several medians), the absolute distance can be conveniently assumed to measure the distance between the ranks of the two alternatives (Kemeny 1959). Young (1995) determined that the absolute distance between alternatives is appropriate if all alternatives are to treated alike; moreover, it gives the same results as Borda's rule. Wittemeyer and Getz (2006) apply the absolute distance to the case named 'bivariate' in this paper.

Therefore, we conjectured that the further we go down the ranking, the more the distance from the first position should more than linearly reduce the weight of the alternative. Formula (4) then changes as follows:

$$W''' = W'''(a = i) = \frac{(A - i + 1)^\delta}{\sum_i^A (A - i + 1)^\delta} \qquad (i = 1, \dots, A; \delta \geq 0) \tag{8}$$

where δ is the power of the distances from the first position. Ahn and Park (2008) call this rule 'rank exponent' (RE) weight. It is easy to see that if $\delta = 0$, all weights are equal, and equal to 0.2, and if $\delta = 1$, Formula (8) coincides with Formula (4), which corresponds to the linearly scaled weights. For $\delta = 2$, the frequencies close to the top position weigh more than linear and diverge from linearity as we approach A, the bottom of the ranking. In fact, in case $A = 5$, the vector of linear weights is [0.333;0.267;0.200;0.133;0.065] while the quadratic vector is [0.454;0.291;0.164;0.073;0.018]. Moreover, the vector of RR weights is [0.438;0.219;0.146;0.109;0.088], the vector of ROC is [0.457;0.257;0.157;0.090;0.040], and the vector of SR is [0.381;0.248;0.178;0.124;0.076].

To weigh a comparison between alternatives a and b, the rules vary according to assumptions. Weights can be computed simply as a function of the distance of alternative a from the first position: if alternative a is in the first position, we assign the largest weight to the comparison between a and j, where j can be any other alternative ($j = b, \dots, A$); if a is in the second position, the weight will be one less than previously, until the next to last position, when the weight will be 1; and for $a = A$, the weight will be zero. The standardised form of weight W_i for a bivariate analysis is a linear function of the distance of the first alternative from the top position in the ranking[3]:

$$W_i'' = W''(a = i) = \frac{A - i}{\sum_i^A A - i} = \frac{2(A - i)}{A(A - 1)} \qquad (i = 1, \dots, A - 1) \tag{9}$$

This weighting system has a maximum of $2/A$ for $i = 1$ and a minimum of $2/(A(A - 1))$ for $i = A - 1$:

$$\frac{2}{A(A - 1)} \leq W'' \leq \frac{2}{A}.$$

It is easily seen that Formula (9) is the same as Formula (4) in cases where i varies between 1 and $(A—1)$. Formulae from (5) to (7) can also be adapted to weigh pair comparisons of alternatives by leaving i to vary between 1 and $(A—1)$ and substituting A with $(A—1)$ in the formulae.

Other possible rules refer either to how many ranks separate the two alternatives or which are the alternatives that occur between the two at stake. For instance, if we compare two alternatives whose positions are $a = 1$ and $b = 4$, should it matter if another alternative c is in the second or the fifth position? From Arrow's theorem (Arrow 1951), we know it does because the rest of the world may in some way influence, even significantly, the relationship between the two compared alternatives. The degree to which it matters and

---

[3] Weighting Formula (8) possesses the following usual distance properties (Cook and Seiford 1978): $d(a, a) = 0; d(a, b) = d(b, a) \geq 0 \ for \ a \neq b = 1, \dots, A$ and $d(a, c) \leq d(a, b) + d(b, c)$, and hence can be called a metric. More detailed axioms can be found in Cook et al. (1986).

in which direction is unknown. In what follows, we model the between-alternative relations according to various hypotheses on the distance between their ranks. We ignore the possible dependence of the magnitude of a dominance relationship on the position of the remaining alternatives.[4] As a matter of fact, we rely on the local independence of the irrelevant alternatives (Young 1995), namely the independence between the two compared alternatives and the other A—2 alternatives not involved in the comparison, an assumption that is reasonable in a case, such as ours, in which the alternatives are social entities just functionally interacting with each other but likely independent in their judgements of the mobility process.

Analogous to the univariate case, we conjectured that the further we go down the ranking, the more the distance from the first position should more than linearly affect the weight of the alternative. So, the distance rule can be powered as:

$$W^{IV} = W^{IV}(a = i) = \frac{(A - i)^{\delta}}{\sum_{i}^{A} (A - i)^{\delta}} \qquad (i = 1, \ldots, A - 1; \ \delta \geq 0) \qquad (10)$$

If $A = 5$, the vector of weights in the linear case is [0.4;0.3;0.2;0.1;0] and in the quadratic case is [0.533;0.300;0.133;0.033;0]. Analogously, the vector of weights is [0.48;0.24;0.16;0.12;0] in the RR weighting system, [0.521;0.271;0.146;0.062;0] in the ROC system, and [0.436;0.273;0.182;0.109;0] in the SR system.

Also, if alternatives a and b are apart in a ranking, we could pose for ourselves the question: *should we equally weight a comparison between a and b if, for instance, a is first and b is second, instead of when a is first and b is in any other non-contiguous position?* Indeed, Bradley and Terry (1952) and Kemeny (1959) suggested to account for how far apart one alternative is from another in a comparison. This question is not addressed in this paper.

2.3 Criteria for Choosing the Best Approach

The literature on how to transform a set of available rankings into a set of scores (as defined in Sect. 2.2) is sparse, and there is no agreed upon solution for comparing these or similar approaches. Therefore, we propose two methods that may help in selecting the best among the proposed methods:

a) *Variability of rankings given by assessors around their aggregate ranking*. This criterion assumes that the aggregate ranking is what the assessors evaluated and that a ranking delivered by an assessor is a trial of the ranking we expected to measure. Therefore, for each group of assessors, ($g = 1, ..., G$), we have n_g rankings whose variability is proportional to the distance between each individual ranking and the group ranking:

---

[4] The possibility to introduce more refined hypotheses about the distance between alternatives in a comparison and the orderings of the alternatives other than those we compare has been the focus of many studies (see David 1987; de Vries 1988; Cook and Kress 1990; Adler, Friedman and Sinuany-Stern 2002; Wang et al. 2007; Llamazares and Peña 2013). The literature on these hypotheses conveys peculiarities that depend on the disciplinary domain, the scope of the analysis, the type and number of the involved alternatives to be ordered or scored (competing animals, agonistic sports teams and competitions, social or business decisions, judgements, etc.), the number of observed rankings, and the methodological drawbacks, thus, leaving many open questions on more refined hypotheses.

$$D_{mg} = \left\{ \frac{1}{n_g} \sum_{k=1}^{n_g} \sum_{a=1}^{A} \left| R_{mgka} - R_{mg0a} \right|^{\delta} \right\}^{1/\delta} \quad (m = 1, ..., M; \ g = 1, ..., G; \ \delta \geq 1) \quad (11)$$

where $|\cdot|$ denotes the absolute distance of the argument; R_{mg0a}, the mean rank of the a th alternative obtained with method $m(m = 1, ..., M)$ at group g; and R_{mgka}, the ranking of the same alternative in the ranking delivered by assessor $k(k = 1, ..., n_g)$ belonging to group g. The difference $R_{mgka} - R_{mg0a}$ measures how many ranks alternative $a(a = 1, ..., A)$ has to move to be in the same position as the group ranking; δ is the power to which the distances between rankings are raised; and all the other symbols have the same meanings as in Sect. 2.2. It can be easily understood that $\delta = 1$ gives the between-rank absolute distance, and $\delta = 2$ gives the Euclidean distance.[5] The former distance is known to smooth the differences between the compared rankings, while the latter one gives much more relevance to large deviations.

By construction, D_{mg} is always positive. It can be standardised to compare the ranking variability from different studies. A possible standardisation is the ratio between the empirical value and its maximum, which is obtained when one half of assessors give rankings in reverse order to the other half of assessors.

In case $\delta = 1$, $\text{Max}(D_{mg}) = A^2/2$ if A is even and $(A^2 - 1)/2$ if A is odd. In case $\delta = 2$, $\text{Max}(D_{mg}) = A(A - 1)(A + 1)/3$. It can be noticed that both maxima are a constant function of A, so they can be used to standardise both single rankings and the aggregate one.

For a group of assessors, the standardised distance, $D_m^* = D_m/\text{Max}\left(D_{mg}\right)$, can be conceived as the proportion of within-ranking positions that disagree with respect to a mean ranking. For instance, a standardised distance evaluated as 0.3 means that, as a whole, 30% of positions in the analysed rankings differ from the reference positions.

The method with the minimum standardised distance has a ranking that best represents the rankings delivered by the n assessors, i.e. the ranking varies the least.

b) *Absolute distance between the final scores generated by a model and the aggregated one mediated over all models.* We assume the aggregated score is that to which all models should conform. In our case, the reference score is the one resulting from the average of the scores estimated for all models. Therefore, the absolute distance between all models and their mean is an estimate of the bias of the models:

$$B_{mg} = \frac{1}{A} \sum_{a=1}^{A} \left| \hat{Y}_{mga} - \overline{\overline{Y}}_{ga} \right| \quad (m = 1, \ldots, M; \ g = 1, \ldots, G) \quad (12)$$

where $\hat{Y}_{mga}$ denotes the estimate of the score for alternative a from the group of assessors n with method m; and $\overline{\overline{Y}}_{ga}$ the average score for alternative a from group g across the compared methods.

---

[5] Fligner and Verducci (1986), instead of a power function of the between-rank distance, adopt an exponential function for which they give mean and standard deviation and a method for significance evaluation.

3 An Application to Mobility Beneficiaries

The models presented in Sect. 2.2 were applied to the rankings delivered by three groups of assessors: (a) the participants in international mobility; (b) the schools and training centres that either sent or hosted participants; and (c) the companies that either sent or hosted participants.[6]

In this application, the alternatives (henceforth, the actors) are the five possible beneficiaries of VET international mobility: participants; schools and training centres; companies (both sending and hosting); the labour market; and the European Union (EU) as an institution. To merge the scores of the assessing groups, the responses obtained from the three groups of assessors were mediated through an (unweighted) arithmetic mean, since, for the purpose of representing the mobility process, the responses given by the three groups are equally informative.

The models are ordered according to computational difficulty:

1. First, we analysed the distribution of the main beneficiaries of international mobility that occurred in the first positions in the assessors' rankings. The frequency gained by actor $a(a = 1, ..., A)$ was the estimate of their benefit score. This model is named Model 1. It could be computed for each group of assessors and, in a mediated form, for all groups of assessors.
2. Then, we computed the linear combination of the frequencies at all rank positions for each mobility actor according to Formula (4), in its standardised form. Formula (4) applies constant weights to the ranking positions. This model, which can be considered a weighted synthesis of each frequency distribution, is named Model 2, and was computed for each group of assessors. Estimates for $\delta = 1$ and $\delta = 2$ and for ROC, RR, and SR rules were computed.
3. We also applied the bivariate analysis to estimate the benefit scores. This application was based on the estimates of the dominance degree of a generic actor a over another generic actor b for all possible combinations of actor pairs. This allowed the construction of a dominance matrix **P** (Formula 3) of which we computed the right eigenvector associated with the first positive eigenvalue (Formula 2). The standardised elements of the eigenvector estimated the benefit scores of the mobility actors. This is presented as Model 3.
4. Finally, we applied the same procedure adopted to compute pair comparisons between actors using weights given in Formula (9). The standardised scores deriving from the first eigenvector of this new matrix are presented as Model 4. Weights were powered both with $\delta = 1$ and $\delta = 2$ and also ROC, RR, and SR rules were applied.

The main results are presented in Table 1, in which the estimates of the benefit scores for the actors of international mobility are compared to each other according to the generating model and the group of assessors.

---

[6] Indeed, schools/training centres and companies could be partitioned according to their duties in the mobility process: some of them were senders of participants, some were hosting units, and others were both sending and hosting units. In this paper, we did not distinguish these sub-categories.

Table 1 Percent estimates of the benefit scores as assessed by participants, by type of model and category of beneficiaries from international mobility *(estimates add up to 100 by row)*.

Models by assessor and weighting power	Beneficiaries of mobility (Alternatives)				
	Participants	Schools and training c.	Companies (send/host)	Labour market	EU as an institution
1—Participants	69.7	11.0	9.3	3.6	6.4
1—Schools	90.7	2.3	2.3	0.0	4.7
1—Companies	75.8	6.4	8.4	2.0	7.4
1—Mean of Assessor Groups	78.7	6.6	6.7	1.9	6.1
2—Participants, RS $\delta = 1$	31.6	18.3	18.5	15.5	16.1
2—Participants, RS $\delta = 2$	43.9	16.3	16.5	11.0	12.3
2—Schools, RS $\delta = 1$	32.8	21.3	18.3	15.1	12.5
2—Schools, RS $\delta = 2$	43.8	20.3	15.8	11.2	8.9
2—Companies, RS $\delta = 1$	30.5	19.1	21.6	15.2	13.6
2—Companies, RS $\delta = 2$	39.5	17.3	21.3	11.5	10.4
2—Assessors Average, RS $\delta = 1$	31.6	19.6	19.5	15.2	14.1
2—Assessors Average, RS $\delta = 2$	42.4	18.0	17.9	11.2	10.5
2—Participants ROC	39.5	16.9	17.0	12.8	13.8
2—Schools ROC	43.8	18.9	15.5	11.7	10.1
2—Companies ROC	39.4	16.9	20.2	12.1	11.4
2—Assessors Average ROC	40.9	17.6	17.5	12.2	11.8
2—Participants RR	32.8	17.9	17.8	15.4	16.1
2—Schools RR	39.8	20.0	16.6	12.0	11.6
2—Companies RR	37.8	16.6	18.9	13.1	13.6
2—Assessors Average RR	36.7	18.2	17.8	13.5	13.8
2—Participants SR	32.2	18.1	18.2	15.4	16.1
2—Schools SR	36.7	19.7	17.0	14.1	12.5
2—Companies SR	33.6	18.0	20.5	14.3	13.6
2—Assessors Average SR	34.1	18.6	18.6	14.6	14.1
3—Participants	33.7	19.0	19.7	13.1	14.5
3—Schools	39.7	20.2	16.7	12.7	10.7
3—Companies	35.0	18.3	21.3	13.3	12.1
3—Average of Assessors	36.1	19.2	19.2	13.1	12.4
3—All assessors, P matrix	35.8	19.2	19.3	13.2	12.5
4—Participants, RS $\delta = 1$	34.4	19.2	19.2	12.7	14.5
4—Participants, RS $\delta = 2$	34.8	19.4	19.0	12.1	14.7
4—Schools, RS $\delta = 1$	39.9	20.2	16.7	12.2	11.0
4—Schools, RS $\delta = 2$	39.9	20.1	16.7	11.7	11.6
4—Companies, RS $\delta = 1$	35.8	18.0	21.2	12.5	12.5
4—Companies, RS $\delta = 2$	36.2	17.9	21.1	11.8	13.0
4—Assessors Average, RS $\delta = 1$	36.7	19.1	19.0	12.5	12.7
4—Assessors Average, RS $\delta = 2$	37.0	19.1	18.9	11.9	13.1
4—Participants ROC	34.7	19.3	19.1	12.3	14.6
4—Schools ROC	39.8	20.1	16.7	11.8	11.6
4—Companies ROC	36.1	17.8	21.0	12.0	13.1

Table 1 (continued)

Models by assessor and weighting power	Beneficiaries of mobility (Alternatives)				
	Participants	Schools and training c.	Companies (send/host)	Labour market	EU as an institution
4—Assessors Average ROC	36.9	19.1	18.9	12.0	13.1
4—Participants RR	34.4	19.2	19.2	12.7	14.5
4—Schools RR	39.8	20.0	16.6	12.0	11.6
4—Companies RR	36.0	17.8	20.9	12.3	13.0
4—Assessors Average RR	36.8	19.0	18.9	12.3	13.0
4—Participants SR	34.4	19.2	19.2	12.7	14.5
4—Schools SR	39.9	20.1	16.7	12.1	11.2
4—Companies SR	35.9	17.9	21.1	12.4	12.7
4—Assessors Average SR	36.7	19.1	19.0	12.4	12.8
Mean Score: Participants[a]	35.1	18.4	18.5	13.2	14.7
Mean Score: Schools[a]	39.6	20.1	16.7	12.4	11.2
Mean score: Companies[a]	36.0	17.8	20.8	12.8	12.6
Mean Score: Overall[a]	36.9	18.8	18.7	12.8	12.8

[a]Results of Model 1 are ignored in computing mean scores

Table 1 shows that estimates differed significantly according to the applied model. The overall mean scores of the alternatives by assessor group and the average of the mean scores are presented at the bottom of Table1 . The results show the following:

- The difference in estimates using just the first positions with respect to the methods based on the full distribution of evaluations is notable: the average of the participants' scores computed over all assessors was almost 80%, while the average of all other methods was 36.9% and the median was 36.7%. This suggests that the estimates based on just the first position of the rankings should be considered a shortcut, as it is not influenced by the relations between beneficiaries. Since the estimation procedure based on just the first positions is an outlier among the applied models, it will not be considered further here.
- Among the other computational methods, Model 2 rules were computationally the least troublesome. The difference between the most extreme score-profiles obtained by adopting the quadratic distance of alternatives and that obtained applying the absolute distance with the same RS rule is relevant: the quadratic strategy produced scores that were, averaging the benefit perceived by all assessors for participants as beneficiaries, more than 10% larger (42.4% instead of 31.6%) than those obtained with the absolute weighting rule. The scores assigned to the other beneficiaries were reduced proportionally to the score assigned to the first beneficiary but were consistent with each other. The score assigned to the top beneficiary according to the ROC rule follows immediately that of the RS with quadratic weighting rule (40.9% on average), then the RR (36.7%) and the SR (34.1) scores. These results are consistent with the literature.
- The weighting rules seemed not so important if scores were estimated through the analysis of dominances. The quadratic procedure yielded results similar to the linearly-weighted analysis of dominances: the mean difference was lower than 0.6%

Table 2 Mean absolute standardised distance between the delivered rankings and the final ranking and bias, by model and group of assessors *(estimates add up to 100 by row; Model 1 excluded).*

Model	Participants		Schools		Companies		Overall	
	Distance	Bias	Distance	Bias	Distance	Bias	Distance	Bias
$2, \delta = 1$	0.538	0.0147	0.294	0.0273	0.337	0.0219	0.498	0.0212
$2, \delta = 2$	0.538	0.0351	0.294	0.0176	0.337	0.0160	0.498	0.0220
2, ROC	0.538	0.0175	0.294	0.0167	0.337	0.0137	0.498	0.0160
2, RR	0.558	0.0142	0.294	0.0023	0.337	0.0124	0.510	0.0065
2, SR	0.538	0.0142	0.294	0.0132	0.378	0.0108	0.498	0.0121
3	0.538	0.0071	0.294	0.0020	0.337	0.0061	0.484	0.0049
$4, \delta = 1$	0.538	0.0059	0.294	0.0017	0.337	0.0024	0.510	0.0027
$4, \delta = 2$	0.538	0.0059	0.294	0.0029	0.378	0.0039	0.510	0.0037
4, ROC	0.558	0.0059	0.294	0.0025	0.378	0.0031	0.510	0.0025
4, RR	0.538	0.0059	0.294	0.0023	0.378	0.0019	0.510	0.0025
4, SR	0.538	0.0059	0.294	0.0013	0.378	0.0018	0.510	0.0027

whichever assessing-group was considered. Furthermore, the differences between assessors using either linear or quadratic powers for weight estimation were less important than those obtained in the univariate analysis. The application of any weighting rule to the analysis of dominances (Model 4) led to score profiles that did not diverge with each other. Actually, the first-beneficiary scores maintained the ordering highlighted in case of univariate analysis (Model 2): the largest value was obtained by the quadratic weighting (37.0% on average), ROC (36.9%), RR (36.8%), SR (36.7%), and RS (36.7%) rules.

- We ascertained just three cases of equal scores between alternatives: once when mediating the scores obtained by three groups of assessors with Model 3 and twice when applying linear weights to Model 4. This may mean that even a flatting rule applied together with the analysis of dominances tends to assign to alternative scores that differ from each other. This result is in line with the literature (e.g.Landau 1951; Hemelrijk, Wantia and Gygax 2005).
- Whatever the weighting rule, there was a clear hierarchy that placed participants at the top of the ranking with an endorsement of approximately 37%, then schools and companies at approximately 19% each, and finally, the labour market and the EU as an institution with approximately 12% each.

Table 2 presents the results of the application of the two statistical criteria suitable for the evaluation of the capability of the models to give adequate estimates. The criteria were a measure of ranking variability, Formula (11), and a measure of bias, Formula (12).

The distance measures in Table 2 refer to the variability of the delivered rankings from the final ranking, computed for each assessing group and model. The mean absolute distances were computed from the values marked with an asterisk at the bottom of Table 1. Moreover, for practical purposes, only complete rankings have been analysed. The bias was computed as the mean absolute deviance from a mean score, computed for each assessing group and for the assessors altogether. The Table 2 estimates allow us to state the following:

- The estimates of the distance between the collected rankings and the reference ranking did not differentiate among models because the reference ranking of the single models was about the same for all models. Thus, the measure of distance for Models 2 and 3 were the same for every group of assessors, but only for the RR rule. Small differences, instead, were found between Models 3 and 4 regardless of the weighting rule. The relative invariance of the distances (no difference exceeded 2.5%) was caused by both the near-linearity of the assessed alternatives and the shortness of the ranking.
- Models 3 and 4, which were based on the analysis of dominances, were able to produce biased estimates of the scores in a lower proportion than Model 2, which was based on just the univariate analysis of frequencies.
- Model 3, which was based on an equal-weight system, performed better in terms of variability than Model 4, for which we introduced various articulated weighting systems, but its bias was higher than any of the weighting systems in Model 4.[7]
- All weighting rules applied to dominance analysis showed the same level of variability. Instead, the bias slightly differed among the applied rules: the ROC and RR rules showed the lowest bias, while the SR and the RS with linear weighting rules showed an 8% higher bias. The highest bias (almost twice that of the ROC rule) was found for the RS rule with quadratic weighting.

4 Discussion and Conclusion

In this paper, we presented a set of rules to score a common profile starting from large sets of rankings. Both rankings and scores concerned five categories of possible recipients of benefits from VET international mobility. For score estimation, we applied various weighting procedures that included at least three levels of information present in the collected rankings: (a) just the top recipient as stated by assessors; (b) the full distribution of frequencies derived from the rankings delivered by the assessors; and (c) the between-alternative dominance relations that can be deduced from the assessed rankings.

We showed that the more information we used in the scoring procedure, the more stable and accurate were the estimates. So, the estimates based on just the first position were shown to be an outlier with respect to the other estimation procedures, and this method would only be recommended for an initial estimation of scores, but not for a definitive estimate. The estimation procedure based on the analysis of univariate frequency distributions, which is intermediate in terms of quantity of processed information, gave estimates that were also intermediate regarding both variability and bias. Models that use the bivariate relations between alternatives on top of the frequency distribution of assessments were the best-performing models. This result is consistent with the literature (see, among others, Poisbleau, Jenouvrier and Fritz 2006), where it has been highlighted that the between-alternative dominance relations takes into account both the identity of the ranked alternatives and the bivariate interactions between alternatives present in the assessor's minds.

---

[7] It is worth saying that the comparison between Models 3 and 4 would have led to the same conclusion had the score estimates been standardised with their sampling error. As Poisbleau, Jenouvrier, and Fritz (2006) also show, the sampling error of estimates based on pair comparisons is a quadratic function of n, the sample size of assessors. In fact, any ranking delivered by an assessor brings about $A(A-1)/2$ distinct pair comparisons, if no ties are allowed, and this makes the sampling errors of the estimated values of the **P** matrix very low with respect to the estimates themselves.

Regarding the optimality of the weighting rules, the comparison between EW, RS, RR, ROC, SR, and RE rules within the processed informative models, showed conflicting results: in both univariate and bivariate analyses the variability criterion highlighted that rules whose configuration is steeper (RR and ROC) perform somewhat worse than or as well as linear or near-linear (RS and SR) rules, while the bias criterion showed the opposite, with RR and ROC slightly better than the linear and quasi-linear one. The quadratic weighting intensifies the steepness of the configuration so much that both the variability and bias criteria highlight its inadequacy as a behavioural model. On the contrary, the equal-weight rule provide a less adequate fit to the basic rankings than any position-based rule. Our data conclusively shows that the first position is to be weighted somewhat more than with a simple linear weighting.

Also, the solution of weighting through an SR model—which is intermediate between the linear and the RR rule—is as adequate as RR in the bivariate analysis and somewhat more adequate than RR in the univariate analysis if we refer to the variability criterion, but is less adequate if the bias criterion is considered. If we adopt the bias criterion, the best rule is RR, while if the variability criterion is adopted, there is no difference between the ROC, RS, and SR rules. In other words, although the between-rule differences are little, the highlights depend on the chosen evaluation criterion.

Our results are in line with the mainstream literature. In fact, Barron and Barrett (1996) compared RS, RR, ROC, and EW rules and stated that the ROC appears to perform better that the others. Sarabando and Dias (2009), applying a variety of rules, found that ROC rule is the best if just the first alternative or the top ranked alternatives are retained. Jia et al. (1998) compared EW, RS, and ROC rules, and also direct rating of alternatives in case the basic rankings are affected by assessment error. They found that direct rating is better if error is small, while RR and ROC rules are better if error is large, but the differences between linear and steeper rules are small. In addition, Jia et al. (1998) found that the accuracy of the rules is sensitive to the number of alternatives and to the sign and level of correlation among basic rankings. In particular, the authors found that if the rankings' fluctuation is high and the number of alternatives is five, the RR rule gives more accurate results than the ROC, and the latter more accurate results than RS. However, the between-rule differences are small. Furthermore, they showed that if the correlation among the rankings is positive and high—as it is in our case—the between-rule differences are even smaller.

So, which weighting method should one use? From our results, the application of a dominance analysis technique gives more accurate estimates than just the univariate analyses. The accuracy of the choice of the weighting rule depends on the steepness of the "true" weights. In summary, EW is one extreme of the horn, that in which weighting is irrelevant because either we purposively ignore the elicited rankings or consider the basic data purely random. On the contrary, the RR, ROC, and RE (with $\delta > 1$) rules highlights that the first alternative is nearly more important than the others and implies steep weight functions. In between, RS and SR rules provide flatter weight functions. Tversky et al. (1988) and Fischer and Hawkins (1993) found that in behavioural analyses and in preference-governed people's choices, functions are often quite steep. This seems true in our case, because in all assessors' minds, participants are those who benefit the most from international mobility, while the EU as an institution and the labour market benefit just indirectly from the scheme and are, consequently, positioned by all assessors at the bottom of the ranking.

In conclusion, our results support the value of surveys on social preferences for elucidating hidden social hierarchies and values orders, and for pinpointing the social categories that are particularly associated with a phenomenon (i.e. those that are often observed at the

top of a ranking, versus others that occur further down the rankings and are only vaguely associated with a phenomenon). In cases where ranks are transformed into scores, these results imply that a more-than-linear weight should be assigned to the top position. That is why the RR and ROC rules applied together within a dominance analysis seem to better fit our mobility data.

References

Adler, N., Friedman, L., & Sinuany-Stern, Z. (2002). Review of ranking methods in the data envelopment analysis context. *European Journal of Operational Research, 140,* 249–265.

Ahn, B. S., & Park, K. S. (2008). Comparing methods for multiattribute decision making with ordinal weights. *Computers and Operations Research, 35,* 1660–1670.

Arrow, K. J. (1951). *Social choice and individual values*. New York: Wiley.

Barron, F. H. (1992). Selecting a best multiattribute alternative with partial information about attribute weights. *Acta Psychologica., 80*(1–3), 91–103.

Barron, F., & Barrett, B. E. (1996). Decision quality using ranked attribute weights. *Management Science, 42,* 1515–1523.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block design 1. *The method of pair comparisons Biometrika, 39,* 324–345.

Brams, S. J., & Fishburn, P. C. (2002). Voting procedures. In K. Arrow, A. Sen, & K. Suzumura (Eds.), *Handbook of social choice and welfare* (pp. 175–236). Amsterdam: Elsevier Science.

Ciavolino, E., & Carpita, M. (2015). The GME estimator for the regression model with a composite indicator as explanatory variable. *Quality and Quantity, 49*(3), 955–965.

Cook, W. D., & Kress, M. (1990). A data envelopment model for aggregating preference rankings. *Management Science, 36,* 1302–1310.

Cook, W. D., Kress, M., & Seiford, L. M. (1986). An axiomatic approach to distance on partial orderings. *Recherche Opérationnelle, 20*(2), 115–122.

Cook, W. D., & Seiford, L. M. (1978). Priority ranking and consensus formation. *Management Science, 24*(16), 1721–1732.

Danielson, M., & Ekenberg, L. (2014). Rank ordering methods for multi-criteria decisions. In: Proceedings of the 14th group decision and negotiation—GDN 2014. Springer

Danielson, M., & Ekenberg, L. (2017). A robustness study of state-of-the-art surrogate weights for MCDM. *Group Decision and Negotiation, 26,* 677–691.

David, H. A. (1987). Ranking from unbalanced paired-comparison data. *Biometrika, 74,* 432–436.

de Borda, J. C. (1784). *Mémoire sur les élections au scrutiny .Histoire de l Académie Royale des Sciences pour*. Paris: Imprimerie Royale.

de Vries, H. (1998). Finding a dominance order most consistent with a linear hierarchy: A new procedure and review. *Animal Behaviour, 55,* 827–843.

Fabbris, L., & Boetti, L. (2019). *ROI-MOB: Measuring the return on investment in VET mobility in the european union*. Padova: Cleup.

Fischer, G. W., & Hawkins, S. (1993). Strategy compatibility, scale compatibility, and the prominence effect. *Journal of Experimental Psychology: Human Perception and Performance, 19,* 580–597.

Fligner, M., & Verducci, J. (1986). Distance based ranking models. *Journal of the Royal Statistical Society B, 48,* 359–369.

Hemelrijk, C. K., Wantia, J., & Gygax, L. (2005). The construction of dominance order: Comparing performance of five methods using an individual-based model. *Behaviour, 142,* 1037–1058.

Honda, N., Tanaka, M., Aida, S., & Tamaki, M. (1981). *Group consensus formation support systems introducing distance between priority rankings* (pp. 1401–1406). Budapest: IFAC 9th Triennial World Congress.

Hudry, O., & Monjardet, B. (2010). Consensus Theories. *Mathematics and Social Sciences, 190*(2), 139–167.

Jensen, R. E. (1986). Comparison of consensus methods for priority ranking problems. *Decision Sciences, 17*(2), 195–211.

Jia, J., Fischer, G. W., & Dyer, J. (1998). Attribute weighting methods and decision quality in the presence of response error: A simulation study. *Journal of Behavioral Decision Making, 11*(2), 85–105.

Kemeny, J. (1959). Mathematics without numbers. *Daedalus, 88,* 575–591.

Landau, H. G. (1951). On dominance relations and the structure of animal societies: I. Effects of inherent characteristics. *Bulletin of Mathematical Biophysics, 13,* 1–19.

Llamazares, B., & Peña, T. (2013). Aggregating preferences rankings with variable weights. *European Journal of Operational Research, 230,* 348–355.

Manisera, M., & Vezzoli, M. (2012). Assessing item contribution on unobservable variables' measures with hierarchical data. *Electronic Journal of Applied Statistical Analysis, 5*(3), 314–319.

Moon, J. W., & Pullman, N. J. (1970). On generalized tournament matrices. *SIAM Review, 12,* 384–399.

Poisbleau, M., Jenouvrier, S., & Fritz, H. (2006). Assessing the reliability of dominance scores for assigning individual ranks in a hierarchy. *Animal Behaviour, 72,* 835–842.

Sarabando, P., & Dias, L. C. (2009). Multi-attribute choice with ordinal information: A comparison of different decision rules. *IEEE Transactions on Systems Man and Cybernetics - Part A Systems and Humans, 39*(3), 545–554.

Stillwell, W. G., Seaver, D. A., & Edwards, W. (1981). A comparison of weight approximation techniques in multiattribute utility decision making. *Organizational Behavior and Human Performance, 28,* 62–77.

Tanino, T. (1988). Fuzzy preference relations in group decision making. In J. Kacprzyk & M. Roubens (Eds.), *Non-conventional preference relations in decision making* (pp. 54–71). Berlin-Heidelberg: Springer-Verlag.

Tversky, A., Sattath, S., & Slovic, P. (1988). Contingent weighting in judgment and choice. *Psychological Review, 95,* 371–384.

Wang, Y. M., Luo, Y., & Hua, Z. (2007). Aggregating preference rankings using OWA operator weights. *Information Sciences, 177,* 3356–4336.

Wittemeyer, G., & Getz, W. M. (2006). A likely ranking interpolation for resolving dominance orders in systems with unknown relationships. *Behaviour, 143,* 909–930.

Young, P. (1995). Optimal voting rules. *Journal of Economic Perspectives, 9*(1), 51–64.

Zanarotti, M. C., & Pagani, L. (2015). Some considerations to carry out a composite indicator for ordinal data. *Electronic Journal of Applied Statistical Analysis, 8*(3), 384–397.