


The Balanced Worth: A Procedure to Evaluate Performance in Terms of Ordered Attributes

Carmen Herrero¹ · Antonio Villar² 

Accepted: 2 December 2017 / Published online: 8 December 2017
© Springer Science+Business Media B.V., part of Springer Nature 2017

Abstract There are many problems in the social sciences that refer to the evaluation of the relative performance of some populations when their members' achievements are described by a distribution of outcomes over a set of ordered categories. A new method for the evaluation of this type of problems is presented here. That method, called balanced worth, offers a cardinal, complete and transitive evaluation that is based on the likelihood of getting better results. The evaluation of each society is based on the probability of obtaining better results with respect to the others. The balanced worth is a refinement of “the worth” (Herrero and Villar in PLoS ONE 8(12):e84784, 2013. <https://doi.org/10.1371/journal.pone.0084784>) that overcomes its excessive sensitivity to the differences, due to the presence of ties. We also discuss how this method can be applied for the case of heterogeneous populations and show how it can be applied in different contexts. An empirical example, regarding life satisfaction in Spain is used to illustrate the working of this method.

Keywords Evaluation method · Categorical variables · Relative group performance

1 Introduction

This paper presents a methodology to evaluate the relative performance of social groups when their achievements are described by distributions of outcomes over an ordered set of categories. This is an abstract framework that encompasses many different situations.

✉ Antonio Villar
avillar@upo.es

Carmen Herrero
cherreroblanco@gmail.com

¹ University of Alicante & Ivie, Alicante, Spain

² Department of Economics, Universidad Pablo de Olavide & Ivie, Ctra. Utrera km. 1, 41013 Seville, Spain

When the information on those achievements consists of numerical variables (e.g. income, expenditures, test scores), the evaluation can be performed applying some standard formulae (inequality indices or conventional summary measures involving the mean, the median or different moments of the corresponding distributions).

There are other cases, however, in which the relevant information about social achievements appears in categorical terms, as it is the case where people report satisfaction levels, educational achievements, or health status perceptions, to mention just three cases. The evaluation of societies in this context is more complex and, in spite of its importance, there are not many robust instruments to properly treat categorical data. We shall focus here on this type of scenario.

1.1 Ordered Categorical Data

Most social evaluation problems involving categorical data correspond to situations in which the different categories can be linearly ordered (i.e. arranged in a ranking from best to worse, say). Take for instance the case in which we are dealing with the analysis of health states. People are usually required to report their perceived health in terms of five categories: *very good*, *good*, *fair*, *bad*, and *very bad*, which are clearly ordered from best to worst. If we consider education levels achieved by the adult population, to take another example, it is common to distinguish amongst *no education*, *primary education*, *secondary education*, and *tertiary education*, now ordered from less to more. Ordered categorical variables naturally appear when we transform numerical variables into categorical ones when those variables are not directly comparable (e.g. using the percentiles of the distribution of the numerical variables).¹ We shall see that the ordered structure of the categorical information helps substantially to deal with this type of evaluation problems.

Let us consider the extremely elementary evaluation problem presented in Table 1, which will serve to illustrate some of the ideas we are going to discuss. It corresponds to the distribution of the populations of two societies, A and B, regarding an attribute that admits three different levels, Good, Fair and Bad (we can think of health states, to fix ideas). That is, 30% of the population in A gets “Good”, 40% “Fair”, and so on.

How can we evaluate the relative performance of those societies?

The most common way of treating ordered categorical data is by introducing some *cardinalisation* in the original information; that is, by giving weights to each of the categories in the appropriate order. For instance, giving the weights 3 for category “Good”, 2 for “Fair” and 1 for “Bad”. Then we can calculate some weighted average of those values (typically a generalised mean) and evaluate those two societies. The problem, needless to say, is that the choice of weights determines the results. So, unless there is some definite basis to set the relative importance of the categories, this evaluation approach may lack robustness or be rather arbitrary.

An alternative way of dealing with the evaluation of ordered categorical data is in terms of *stochastic dominance*. If we want to assess the relative goodness of two distributions, we may consider the comparison of their respective cumulative information. Going back to the

¹ This happens, for instance, when analysing the citation impact of research articles regarding different disciplines. As the number of citations is rather idiosyncratic, one usually takes the shares of the papers in the different percentiles of each discipline to make an analysis independent of the differences in the mean and the variance of the citations distributions.

Table 1 Example 1

Categories	Good	Fair	Bad
Population A	0.3	0.4	0.3
Population B	0.4	0.35	0.25

Table 2 Example 1

Categories (cum)	Good	Fair or Good	Fair, Good or Bad
Population A	0.3	0.7	1
Population B	0.4	0.75	1

Table 3 Example 2

Categories	Good	Fair	Bad
Population A	0.3	0.4	0.3
Population C	0.4	0.25	0.35

example in Table 1, let us construct the corresponding cumulative distributions, presented in Table 2.

Those cumulative distributions tell us that Population B is better than Population A, because the probability of a generic individual in B to be “Good” is higher than the same probability in population A, and the probability of being “Good or Fair” is also higher in B than in A. We say that B stochastically dominates A and, given the context, that means that B is *better* than A.

Previous procedure is indeed robust, and only relies on the categorical information without any reference to external weights. It presents, nonetheless, two significant drawbacks. First and foremost, it only provides a partial order, as not all pairs of distributions can be compared this way. This would happen, for instance, if population B had 25% in category Fair and 35% in category Bad. That is, in many cases we cannot say whether a distribution is better or worse than another one according to this evaluation criterion. Second, even in the case in which we can assert that one distribution is better than another, we cannot say how much better, because stochastic dominance does not provide cardinal information about the relative goodness of the distributions.

Liebersson (1976) provides a cardinal evaluation protocol that can be applied to any pair of distributions (a *complete* criterion for this domain of problems), with a slightly different perspective. Rather than comparing cumulative distributions he calculates the total ex-ante probability of a representative agent being better off in one distribution than in the other (called *domination probabilities*). That is, how likely is that an agent from A chosen at random will belong to a higher category than an agent from B chosen at random. The higher that probability, the better. Let us illustrate this idea by considering example 2 in Table 3, which is a variant of example 1.

Populations A and C cannot be compared in terms of stochastic dominance, as mentioned above. As categories are ordered from best to worst, we may compute the total ex-ante probability of a generic individual in A to be better off than an individual in C, p_{AC} , as follows:

Table 4 Example 3

Categories	I	II	III	IV
Population 1	0.1	0.3	0.6	0
Population 2	0.5	0	0	0.5
Population 3	0	0.7	0.2	0.1

$$p_{AC} = 0.3(0.25 + 0.35) + 0.4 \times 0.35 = 0.32$$

Similarly, we may compute the probability of an individual in *C* to be better off than any of *A*, p_{CA} , as:

$$p_{CA} = 0.4(0.4 + 0.3) + 0.25 \times 0.3 = 0.355$$

We can conclude, therefore, that *C* is better than *A*, because this society offers better chances to a generic individual. This method can be applied successfully to any pair of distributions and always provides a definite ranking. It also provides a measure of how much better is one distribution than the other (in the example, *C* is slightly better than *A* as the two domination probabilities are pretty close).

This evaluation procedure, however, turns out to be non-transitive when applied to more than two populations, so that it is not applicable in a general context because we may find cycles and thus be unable to rank distributions. The example in Table 4, involving three groups, 1, 2 and 3, and four categories illustrates this case.

Here, we have that $p_{12} = 0.5$; $p_{21} = 0.45$; $p_{23} = 0.5$; $p_{32} = 0.45$; $p_{13} = 0.25$; $p_{31} = 0.42$, that is, group 1 is better than group 2, group 2 is better than group 3, but group 3 is better than group 1.

Herrero and Villar (2013) present a transitive extension of Lieberman’s idea to more than two populations. They do so by obtaining a vector of evaluations, called the *worth*, that enables both to order any set of populations in terms of their relative goodness and also to provide a cardinal measure of their relative desirability. The basic insight is the following: when comparing any two societies in a context in which there are three or more, it is not enough to compute the relative domination probabilities between those two societies. One has to keep track of how those two societies relate to the rest as well. The bottom line is that dominating a distribution that dominates many others is more important than dominating another that does not fare very well with respect to the rest. This is an idea that appears in many contexts, as it is the case in the analysis of citation impact where citations are pondered by the relevance of the citing journals (Palacios-Huerta and Volij 2004), or the way of ranking pages by Google (Slutzki and Volij 2006).

It can be shown that the worth vector always exists, is generically unique and can be easily computable as it corresponds to the dominant eigenvector of a Perron matrix suitably built. Each component of the vector consists of a ratio of the weighted domination probabilities of a group vis a vis the rest (with weights given by the corresponding worth values), and the sum of the probabilities of being dominated by some other group. In the case of an evaluation involving just two populations the worth vector, $\mathbf{v} = (v_1, v_2)$ is determined by the following intuitive formula:

$$\frac{v_1}{v_2} = \frac{p_{12}}{p_{21}}$$

(i.e. the worth values are proportional to the corresponding domination probabilities).

1.2 The Balanced Worth

This paper introduces a refinement of the *worth* that we call the *balanced worth*, which overcomes its excessive sensitivity to the differences in the outcome distributions, by taking into account the probability of ties. It also provides an extension to the case of heterogeneous populations and a self-contained explanation of the method and its properties.

To understand the need of such a refinement let us focus on the evaluation of two societies. The formula of the worth presented above consists of the ratio of the domination probabilities between the two societies. This implies that the evaluation concentrates on the parts of the distributions in which the groups differ and ignores that in which they are similar. As a consequence, the worth may strongly overestimate those differences when the probability of ties is relevant. The following example illustrates this feature. Suppose we are comparing two groups, i and j , whose distributions yield the following values for the corresponding domination probabilities: $p_{ij} = 0.002$, $p_{ji} = 0.001$. The worth produces the following evaluation: $v_i/v_j = 2$. That is, distribution i is regarded as twice as good as distribution j . Yet if one computes the probability of getting agents within the same level of achievement, we find that this has probability $1 - p_{ij} - p_{ji} = 0.997$. This strongly suggests that both distributions are practically identical and hence that the worth overestimates the relative goodness of distribution i . The balanced worth, which here consists of adding to the domination probabilities the probability of ties equally split between both groups, yields the evaluation $w_i/w_j = 1.002$, which is much closer to what the intuition suggests.²

The balanced worth is described formally in Sect. 2, providing the rationale of this evaluation method, presenting its main properties, and including an empirical illustration regarding life satisfaction in Spain by age groups. Section 3 extends the evaluation method in a different direction, by considering its applicability to heterogeneous populations and provides a method of disentangling the differences in achievements from those derived from differences in the composition of the population being compared. The empirical illustration regarding life satisfaction is discussed in this new context by considering the differences by gender within the age groups. A few final comments in Sect. 4 conclude.

1.3 References to the Literature

In the case of two populations, the seminal paper is Lieberman (1976) who introduces the ideas of probability dominance and its use to compare categorical distributions. These ideas appear also in Cuhadaroglu (2013) in the analysis of discrimination. Related evaluation criteria appear in the statistical measure of distributional similarities (Li et al. 2009; Martínez-Mekler et al. 2009; Gonzalez-Diaz et al. 2014), the ranking of income distributions (Shorrocks 1983; Bellù and Liberati 2005; Bourguignon et al. 2008; Yalonetzky 2012; Sheriff and Maguire 2013), the analysis of segregation and discrimination (Reardon and Firebaugh 2002; Grannis 2002; Echenique and Fryer 2005; Chakravarty and Silber 2007; Frankel and Volij 2011).

When considering more than two groups, Herrero and Villar (2013) present the first extension of Lieberman (1976). Related ideas, for more than two groups also appear in the evaluation of scientific influence (Pinski and Narin 1976; Laband and Piette 1994;

² Note that, for a given problem, the probability of ties will depend on the number of admissible categories defined. The difference between the balanced worth and the worth will thus be smaller the finer the grid of possible outcomes and vanishes for continuous distributions.

Palacios-Huerta and Volij 2004; Crespo et al. 2013), the comparison of network structures (Rosvall and Bergstrom 2007), or the allocation of scores in tournaments (Laslier 1997; Slutzki and Volij 2006).

Besides the three applications provided in the original paper by Herrero and Villar (2013), the worth has been used in a number of studies to evaluate different problems. Herrero et al. (2014) analyse the evaluation of scholastic performance using PISA data and applying inverse probability weighting (IPW) techniques to control for differences in the distribution of the determinants of the outcome variable. Villar (2014) deals with the study of the results of the Programme for International Assessment of Adult Competences (PIAAC), regarding Spain, in the field of *Mathematics*. The key element consists of comparing the relative skills acquired by the different generations that compose the Spanish working age population. The study uses the distributions of the population of the different cohorts into the five proficiency levels defined by the OECD. Gallén and Peraita (2015) provide an application of the worth to the analysis of corporate social responsibility (CSR) engagement in the OECD. The interest of this question derives from the observed expansion of CSR engagement of the OECD countries in recent years, a period of financial crisis. Torregrosa (2015) uses the worth to analyse the evolution of autonomic-nationalist feelings in Spain based on opinion surveys regarding the state of Spanish Autonomous Communities carried out by Spain's Centre for Sociological Research since 1996. Albarrán et al. (2017) analyse the intellectual influence by countries and research fields, from a dataset consisting of 4.4 million research articles published in the period 1998–2003 and indexed by Thomson Scientific, as well as the citations they received during a 5-year citation window for each year in that period. Different conventional evaluation criteria are considered and confronted with the worth.

2 The Model

2.1 The Reference Problem and the Evaluation Method

The reference problem consists of evaluating the relative performance of a collection of g populations, $G = \{1, 2, \dots, g\}$, whose achievements are described by a distribution of values over a finite set of categories that are linearly ordered (ordinal categorical variables). Those populations, also called *groups*, are to be understood as related in some way, e.g. they correspond to subsets of a larger set, such as the plants of a firm, the regions of a country or the countries of a Federation. This is so as making a relative comparison otherwise makes little sense.

Each population $i \in G$ has n_i elements, also called *members*. There is a value associated with each element that measures individual performance, referred to as *outcomes*, which we assume can take on a finite number of values, called *categories* or *levels*. We assume that those levels are ordered from best to worse. That is, level 1 is better than level 2, level 2 is better than level 3, etc.

A *distribution* of outcomes for population i is a vector $\mathbf{a}(i) = (a_{i1}, a_{i2}, \dots, a_{is})$ that describes the fraction of its members into each admissible level of performance. That is, $a_{ir} = n_{ir}/n_i$, where n_{ir} is the number of elements in population i with outcome level r . Clearly, $a_{ir} \geq 0$, $\sum_{r=1}^s a_{ir} = 1$.

An *evaluation problem*, or simply a *problem*, refers to the comparison of the relative performance of those populations in terms of the behaviour of their members. That is,

assessing the relative goodness of the distributions $\mathbf{a}(1), \mathbf{a}(2), \dots, \mathbf{a}(g)$. An evaluation problem can thus be summarized by the matrix \mathbf{A} comprising all those $\mathbf{a}(i)$ distributions, which we interpret as the rows of \mathbf{A} .

The basic principle of comparing the populations' performance refers to the probability of getting better outcomes. For a given problem \mathbf{A} we denote by p_{ij} the probability of a member chosen at random from population i exhibiting a higher level of performance than a member chosen at random from population j . As the levels are ordered from best to worst, we can calculate that probability as follows (see Lieberman (1976):

$$p_{ij} = a_{i1}(a_{j2} + \dots + a_{js}) + a_{i2}(a_{j3} + \dots + a_{js}) + \dots + a_{i(s-1)}a_{js}$$

Let $e_{ij} = e_{ji}$ stand for the probability of a member of group i exhibiting the same level of performance than a member of group j .

$$e_{ij} = a_{i1}a_{j1} + \dots + a_{is}a_{js}$$

By construction, we have: $1 = p_{ij} + p_{ji} + e_{ij}$.

A procedure to obtain quantitative estimates of the relative desirability of those distributions of ordered categorical data is now described. This procedure can be seen in terms of a contest in which each group is confronted randomly with another.

2.1.1 The Simplest Case: Two Groups

Suppose we have just two groups, i and j . In order to determine which group exhibits a better distribution of outcomes, we propose the following protocol. One member from each group will be selected at random and they will be confronted. If the member from group i beats that from group j (that is, it exhibits a higher level of performance), then the distribution of group i is declared *better than* that of group j , and vice versa. Should both members exhibit the same level of performance, a coin is tossed and each group will be declared best with probability $\frac{1}{2}$.

The probability of group i being declared better than group j is given by:

$$p_{ij} + (e_{ij}/2)$$

That is, the probability of i beating j plus half the probability of a tie. Similarly, the probability that group j be the winner in this confrontation is:

$$p_{ji} + (e_{ji}/2)$$

Given these data, how should we value the outcomes of those two groups? Our proposal is simple and natural: *make the value of each group proportional to the probability of being a winner*. That is, if we call w_i, w_j the evaluations of those two groups, we have:

$$\frac{w_i}{w_j} = \frac{p_{ij} + (e_{ij}/2)}{p_{ji} + (e_{ij}/2)} \tag{1}$$

We refer to this evaluation principle as *proportionality*. Note that this formula has one degree of freedom, so that we can choose units arbitrarily. For the case of two groups, therefore, the proportionality principle fully determines the evaluation formula, except for the choice of units.

Equation (1) can be rewritten as:

$$w_i = \frac{[p_{ij} + (e_{ij}/2)]w_j}{p_{ji} + (e_{ji}/2)} \tag{1'}$$

In this way the evaluation of group i appears as the ratio of two interesting expressions. The one in the numerator can be regarded as the *relative advantage* of i over j , as it corresponds to the probability of getting better outcomes, weighted by the evaluation of group j . The denominator can be seen as the *relative disadvantage* of group i with respect to population j , as it expresses the probability of getting worse outcomes.

2.1.2 The General Case: $g \geq 2$ Groups

It is easy to check that if we apply this criterion for pair-wise comparisons when there are more than two groups, we may find a cycle, as the evaluation they induce is not transitive. The example in Table 4 above also serves to illustrate this problem. Now we find that: (1) $p_{12} + (e_{12}/2) = 0.525$, $p_{21} + (e_{21}/2) = 0.475$, which implies that group 1 is better than group 2. (2) $p_{23} + (e_{23}/2) = 0.525$, $p_{32} + (e_{32}/2) = 0.475$, which implies that group 2 is better than group 3. And (3) $p_{31} + (e_{31}/2) = 0.585$, $p_{13} + (e_{13}/2) = 0.415$, which implies that group 3 is better than group 1, thus creating a cycle.

The simplest way of extending the evaluation for more than two groups avoiding previous problem is by taking expectations. That is, the value of group i will be given by the following formula:

$$w_i = \frac{\frac{1}{g-1} \sum_{j \neq i} (p_{ij} + (e_{ij}/2))w_j}{\frac{1}{g-1} \sum_{j \neq i} (p_{ji} + (e_{ji}/2))}, \quad i, j = 1, 2, \dots, g \tag{2}$$

This expression is a generalization of Eq. (1'). The numerator now describes the *average relative advantage* of the distribution of population i with respect to the rest, whereas the denominator corresponds to the *average relative disadvantage* of population i with respect to the rest. Trivially, Eq. (2) collapses to Eq. (1') when there are only two populations.

2.1.3 Balanced Worth

Equation (2) provides a valuation vector $\mathbf{w} = (w_1, \dots, w_g)$ where the value attached to each group is related to the valuation of any other group. In this way, Eq. (2) can be interpreted as an extension of the proportionality principle in Eq. (1) to more than two groups. Note that Eq. (2) does not provide directly the components of the valuation vector, since it requires solving the following system of simultaneous equations.

$$\left. \begin{aligned} w_1 \sum_{j \neq 1} [p_{j1} + (e_{j1}/2)] &= \sum_{j \neq 1} [p_{1j} + (e_{1j}/2)] w_j \\ w_2 \sum_{j \neq 2} [p_{j2} + (e_{j2}/2)] &= \sum_{j \neq 2} [p_{2j} + (e_{2j}/2)] w_j \\ \dots & \\ w_g \sum_{j \neq g} [p_{jg} + (e_{jg}/2)] &= \sum_{j \neq g} [p_{gj} + (e_{gj}/2)] w_j \end{aligned} \right\} \tag{3}$$

The solution to this system we call *balanced worth vector*, as it is a refinement of the concept of *worth* introduced in Herrero and Villar (2013). Theorem 1 in the “Appendix” proves the existence, positiveness and uniqueness (under general conditions) of this vector.

The main properties of the balanced worth vector can be summarised as follows:

- The balanced worth vector, (w_1, \dots, w_g) always exists and it is unique except for the choice of units (it has one degree of freedom).
- This vector provides two types of information: (1) it orders the desirability of the groups, in the sense that higher values correspond to better groups, and (2) it provides a cardinal evaluation of the *relative* goodness of the groups, as the quotient w_k/w_j measures how much better group k is with respect to group j .
- The balanced worth provides a *relative evaluation* of the different groups, as each individual value depends on the data of all the groups. In particular, the balanced worth of a group cannot be computed in isolation.
- The balanced worth attaches to each group the ratio between the average relative advantage of that group and the average relative disadvantage. It is, therefore, a rather intuitive evaluation procedure.
- The balanced worth is *anonymous*, that is, the evaluation only depends on the groups’ performance and not on other aspects such as labels or names. Therefore, permuting the realizations between the groups will not change the evaluation.
- The balanced worth is *symmetric*, namely, if two groups have identical distributions, then the corresponding components of the balanced worth vector are identical.
- The balanced worth is *monotonic*. This means that if group j improves their outcomes whereas all other groups’ outcomes remain unaltered (that is, the distribution of group j shifts to the upper levels of performance), then the balanced worth of group j will (relatively) increase. This property implies *stochastic dominance*: If the distribution of one group stochastically dominates the distribution of another, then it will exhibit a larger balanced worth component.
- The balanced worth of a fully dominated group is zero.³ Moreover, the relative values of the remaining groups do not change if we delete the fully dominated group from the problem.

The computation of the balanced worth can be directly obtained through a friendly and freely available algorithm, hosted on the website of the Instituto Valenciano de Investigaciones Económicas (Ivie) at <http://www.ivie.es/balanced-worth/>. This webpage explains how this algorithm works (it computes the dominant eigenvector of a suitable Perron matrix) and how to proceed to implement the calculations. In particular, the balanced worth can be obtained directly from the matrix of relative frequencies that can be plugged into the algorithm as an Excel table, thus saving much time and effort. By default the algorithm normalizes the values of the balanced worth making the mean of the groups equal to 1.

³ We say that group k is fully dominated when $p_{kj} = 0, \forall j \neq k$.

2.2 An Empirical Illustration: Life Satisfaction in Spain

Let us illustrate the working of this evaluation method by considering the problem of assessing life satisfaction in Spain.

During 2013, the European Union (EU) first elaborated a comparative study regarding the quality of life in the Member States, from a subjective perspective (see Eurostat 2015). The data were collected through the 2013 EU SILC ad-hoc module on subjective wellbeing. Life satisfaction is one of the three dimensions that define subjective wellbeing, based on an overall cognitive assessment (the other two being *affects* and *eudaimonics*). Life satisfaction represents how a respondent evaluates life as a whole, that is, an assessment comprising all areas of a person's existence. It focuses on how people are feeling "these days" rather than specifying a longer or shorter time period (see Veenhoven 1991, 3; Pavot and Diener 2008, 137). Economists may think of that as a measure of individual welfare.

Life satisfaction is measured on a 0–10 scale (where 0 is "not satisfied at all" and 10 "fully satisfied"). To facilitate analyses, those numerical evaluations were grouped into different categories, according to the statistical distribution of the answers. In the case of Spain, the National Statistical Office (INE) used four categories that we term: Low (0–4 points), Fair (5–6 points), High (7–8 points) and Very high (9–10 points). Table 5 provides the distribution of answers by different age groups, together with the balanced worth (normalised so that the mean equals 1) and the normalised means of the different age groups (global mean equals 1).

The most obvious message of these data is that life satisfaction diminishes with age. More interesting is the comparison between the balanced worth, which computes the differences in the distributions by age groups, and the (normalized) means. Even though both measures exhibit a decreasing pattern with age, the differences by age groups are much larger in the case of the balanced worth. Indeed, the coefficient of variation of the balanced worth values is almost four times that of the mean values.

3 Heterogeneous Populations

An implicit assumption of the evaluation model described above is that groups are homogeneous, so that the distribution of the outcome variable is the sole relevant information. Yet, when groups are heterogeneous, one might be interested in evaluating not only the observed outcomes, but also the extent to which those outcomes reflect diverse structural characteristics of the groups that affect the agents' performance.

Aspects such as sex, race, age, nationality, parental background, or wealth, can influence individual outcomes in particular problems and it is interesting to know to what extent the observed outcome differences correspond to differences in the composition of the groups.

There is a number of related but different questions that can be addressed when dealing with heterogeneous populations and the use of balanced worth has to be adapted to each case. Think, for instance, we are evaluating perceived health in OECD countries. Each individual in the sample rates her perceived health in one out of five different health states, ranging from very good to very bad. If we identify each OECD country with a group and apply the balanced worth to evaluate the health state of those countries, we are disregarding the fact that part of the observed differences in the distribution of responses reflects differences in the demographic composition of the populations (there is strong evidence that health perceptions are age dependent).

Table 5 Life satisfaction in Spain by age groups (2013). *Source:* Instituto Nacional de Estadística. Módulo 2013 Encuesta de Condiciones de Vida

	Average satisfaction					Very high 9–10	Balanced worth	Normalized means
	Low 0–4	Fair 5–6	High 7–8	Very high 9–10	Balanced worth	Normalized means		
Total	6.9	9.7	45.2	18.4				
From 16 to 29 years	7.3	6.6	49.1	24.0	1.2283	1.0580		
From 30 to 44 years	7.0	8.9	45.9	20.5	1.0481	1.0145		
From 45 to 64 years	6.7	10.4	45.8	14.5	0.8786	0.9710		
65 years or over	6.6	12.6	39.8	16.8	0.845	0.9565		
Coef. variation					0.153	0.04		

Mean values, distribution by categories (%), and balanced worth

How to address this problem mostly depends on the type of comparisons deemed relevant. One possibility is considering each population subgroup as a different group, so that the evaluation is made with respect to the $\tau \times g$ subgroups, where τ is the number of different types within each group. We call this the *joint evaluation*. In the example of the health states, this means that we think relevant comparing the health status of young people in France relative to old people in Germany, to give an example. Another possibility is that of making comparisons among population subgroups with similar characteristics (e.g. the health states of the young in all countries). We shall refer to those comparisons as the *separate evaluation by types*. It provides an evaluation of the *between groups* relative performance by types. Still a different evaluation problem in the context of heterogeneous populations refers to the evaluation of the degree of heterogeneity within the groups. In the health example, that amounts to evaluate how different are the results on perceived health between generations in different countries. We call this *separate evaluation by groups*. This evaluation provides a measure of *within group* heterogeneity.

Which form of comparison is more adequate depends on the problem at hand and it is part of the modelling choices open to the researcher. We shall now describe briefly how to deal with those questions.

3.1 Separate Evaluation by Types

The evaluation problem in the case of heterogeneous populations can be framed as follows. We have, as before, an evaluation problem involving g groups whose achievements regarding some aspect are given in terms of s ordered levels. The novelty now is that the population of each of those g groups can be classified in terms of τ different *types*, indexed by $t = 1, 2, \dots, \tau$. Each type within a group gathers those members with similar characteristics, so that different types correspond to differential structural traits in the population of that group. In the example regarding perceived health the types are usually defined by age intervals (e.g. young, adult and old), so that the implicit assumption is that all agents in the same age interval are directly comparable in terms of their health states.

The outcome of each group $i = 1, 2, \dots, g$ will now be described by a collection of τ distributions, $\mathbf{a}^t(i) = (a^t_{i1}, a^t_{i2}, \dots, a^t_{is})$, for $t = 1, 2, \dots, \tau$ (a contingency table). Each term $a^t_{ir} = \frac{n^t_{ir}}{n^t_i}$ corresponds to the share of the population of type t within group i with level of achievement r . Here n^t_{ir}, n^t_i are the number of members of group type t with level r within group i , and the total number of members of type t within group i , respectively. For all $t = 1, 2, \dots, \tau$, all $i = 1, 2, \dots, g$, we have: $\sum_{r=1}^s a^t_{ir} = 1$.

We can now evaluate the relative performance of each type among the groups (e.g. comparing health states between old people across countries), by considering the evaluation problem defined by the following collection of $(g \times s)$ -matrices:

$$\mathbf{A}(t) = \begin{bmatrix} \mathbf{a}^t(1) \\ \mathbf{a}^t(2) \\ \dots \\ \mathbf{a}^t(g) \end{bmatrix}, \quad t = 1, 2, \dots, \tau \tag{4}$$

The balanced worth of each of those problems, $\mathbf{w}(t)$, $t = 1, 2, \dots, \tau$, tells us about the relative performance of the corresponding type across groups. The implicit assumption is that comparing the outcomes of different types is not relevant.

The overall evaluation of the group can be obtained as a weighted average of those types, with weights corresponding to the population shares. That is,

$$W_i(\mathbf{A}, \tau) = \sum_{t=1}^{\tau} \frac{n_i^t}{n_i} w_i(t) \tag{5}$$

Each term of this sum in Eq. (5) is the product of two numbers. The first one is the share of type t in the group and reflects its *composition*. The second evaluates the performance of type t in this group relative to type t members of other groups. It provides a measure of the *return* of the type in this group, relative to the return of the same type in other groups.

We can now estimate the composition effect by comparing that value in Eq. (5) with one in which the composition of group i corresponds to a given standard, $W^C(\cdot)$. Suppose that we take the average composition of the groups as the standard, for the sake of simplicity. That yields,

$$W_i^C(\mathbf{A}, \tau) = \sum_{t=1}^{\tau} \frac{\sum_{i=1}^g n_i^t}{\sum_{i=1}^g n_i} w_i(t)$$

The composition effect will thus be measured by:

$$C(\mathbf{A}, \tau) = W_i(\mathbf{A}, \tau) - W_i^C(\mathbf{A}, \tau) = \sum_{t=1}^{\tau} \left(\frac{n_i^t}{n_i} - \frac{\sum_{i=1}^g n_i^t}{\sum_{i=1}^g n_i} \right) w_i(t) \tag{6}$$

Similarly, we may be willing to calculate the effect of the differential returns of the types by comparing (4) with some standard. If we choose the average return, we would have:

$$W_i^R(\mathbf{A}, \tau) = \sum_{t=1}^{\tau} \frac{n_i^t}{n_i} \left(\frac{1}{g} \sum_{i=1}^g w_i(t) \right)$$

Note that, according to our default normalization, the average balanced worth is set equal to one, so that we have:

$$R_i(\mathbf{A}, \tau) = W_i(\mathbf{A}, \tau) - W_i^R(\mathbf{A}, \tau) = \left[\sum_{t=1}^{\tau} \frac{n_i^t}{n_i} w_i(t) \right] - 1 \tag{7}$$

3.2 Separate Evaluation by Groups

A different problem regarding heterogeneous populations is that of measuring the relative performance of the different types *within* groups and providing a summary measure of their degree of diversity.

Assume, as before, that each group $i = 1, 2, \dots, g$ consists of τ different types. The outcome distribution of group i will be given by a matrix:

$$\mathbf{A}(i) = [\mathbf{a}^1(i), \mathbf{a}^2(i), \dots, \mathbf{a}^{\tau}(i)] \quad i = 1, 2, \dots, g$$

where $\mathbf{a}^t(i) = (a_{i1}^t, a_{i2}^t, \dots, a_{is}^t)$, for $t = 1, 2, \dots, \tau$, is the vector that describes the shares of type t within group i into the different levels of achievement.

The balanced worth of each of those partitioned groups, considered in isolation, $\mathbf{w}(i) = (w_1(i), w_2(i), \dots, w_\tau(i))$, for $i = 1, 2, \dots, g$, tells us about the relative performance of the types within group i . Depending on the problem under consideration and the nature of the types, those values may provide measures of segregation, discrimination, intergenerational progress, etc.

A real-valued measure of the degree of heterogeneity for group i can be obtained from the dispersion of those values. Such a measure would permit one comparing heterogeneity between groups, in terms of the dispersion of the components of the balanced worth of their constituent types. Two remarks are to be made on this regard. First, we have to take into account the differences in size of the types when defining this overall heterogeneity measure. Second, the appropriate dispersion measure may vary depending on the problem under consideration (in particular on whether we want to attach differential weights to the relative achievements of the different types).⁴

3.3 The Joint Evaluation

In some cases we might be willing to perform a joint evaluation. That is, comparing all types of all groups as if they were different populations. In this case we would simply apply the balanced worth, $\mathbf{w}(\mathbf{A}, \tau) = [(w_{11}, \dots, w_{1\tau}), \dots, (w_{g1}, \dots, w_{g\tau})]$, to the extended problem consisting of $g \times \tau$ sub-groups. Out of this evaluation we could recover the evaluation of the groups in terms of a weighted sum, with weights corresponding to the population shares. That is,

$$w_i(\mathbf{A}, \tau) = \sum_{t=1}^{\tau} \frac{n_{it}}{n_i} w_{it} \tag{8}$$

Note that the evaluation of group i in Eq. (8) may differ from that obtained in Eq. (5), even though both are weighted sums of group i 's types values. And it may also be different from the within group evaluation, $w_i(i)$. The reason is that $w_i(t) \neq w_{it} \neq w_i(i)$ because each evaluation provides a relative measure of goodness of type t of group i with respect to different terms of comparison. The value $w_i(t)$ is the relative evaluation of type t from group i with respect to type t populations of other groups. The value w_{it} , on the contrary, is the relative evaluation of type t from group i with respect to all other types no matter the groups they belong to. Finally, the value $w_i(i)$ corresponds to the relative evaluation of type t from group i with respect to all other types within this group.

We can also derive an overall evaluation of the types, given by:

$$w_t(\mathbf{A}, \tau) = \sum_{i=1}^g \frac{n_{it}}{n_t} w_{it} \tag{9}$$

This evaluation will also differ from the one obtained by averaging the $w_i(i)$ values, for the same reason explained above.

⁴ Let us recall here that inequality measures typically give more weight to the realizations in the lower part of the distribution. This makes sense when heterogeneity is bad but this is not always the case. For instance when comparing years of schooling across generations in a given country, one would typically like to find that the young generation has higher values than the old one, so that perfect equality is not the desideratum.

Table 6 Life satisfaction in Spain (2013) by age groups and gender. *Source:* Instituto Nacional de Estadística. Módulo 2013 Encuesta de Condiciones de Vida

	Average satisfaction	Low 0–4	Fair 5–6	High 7–8	Very high 9–10
Men					
Total	6.9	9.9	26.7	45.2	18.2
From 16 to 29 years	7.2	7.8	20.3	47.6	24.2
From 30 to 44 years	6.9	9.7	26.3	44.5	19.6
From 45 to 64 years	6.7	10.9	29.0	46.1	14.0
65 years or over	6.8	10.7	29.8	42.4	17.1
Women					
Total	6.9	9.6	26.6	45.1	18.7
From 16 to 29 years	7.3	5.4	20.3	50.5	23.8
From 30 to 44 years	7.1	8.0	23.2	47.4	21.4
From 45 to 64 years	6.7	10.0	29.7	45.4	14.9
65 years or over	6.5	14.0	31.6	37.9	16.6

Table 7 Life satisfaction in Spain (2013). Separate evaluation by types

	BW Men	Norm. mean Men	BW Women	Norm. mean Women
From 16 to 29 years	1.2114	1.0435	1.2408	1.0580
From 30 to 44 years	1.0057	1.0000	1.0898	1.0290
From 45 to 64 years	0.8768	0.9710	0.8727	0.9710
65 years or over	0.9061	0.9855	0.7967	0.9420
Coef. variation	0.131	0.027	0.176	0.046

Table 8 Life satisfaction in Spain (2013). Separate evaluation by age groups

	Balanced worth	Normalized mean
From 16 to 29 years		
Men	0.9798	0.9931
Women	1.0202	1.0069
From 30 to 44 years		
Men	0.9485	0.9857
Women	1.0515	1.0143
From 45 to 64 years		
Men	0.9895	1.0000
Women	1.0105	1.0000
65 years and over		
Men	1.0529	1.0226
Women	0.9471	0.9774

Table 9 Life satisfaction in Spain (2013). Joint evaluation

	Balanced worth	Normalized mean values
16–29 Men	1.1975	1.0435
30–44 Men	0.9936	1.0000
45–65 Men	0.8655	0.9710
> 65 Men	0.8948	0.9855
16–29 Women	1.2553	1.0580
30–44 Women	1.1027	1.0290
45–65 Women	0.8839	0.9710
> 65 Women	0.8067	0.9420
Coef. variation	0.156	0.038

3.4 Life Satisfaction in Spain Revisited

Let us consider now that life satisfaction is gender dependent, thus enriching the empirical example in Sect. 2.2. We keep age groups as our reference groups and consider two different types within each of those groups, men and women. Table 6 provides the basic data.

Consider now the separate evaluation by types. We aim at assessing how life satisfaction varies among men by age groups, and how life satisfaction varies among women by age groups. Table 7 provides the results in terms of the balanced worth and the (normalised) mean values. Mind that, even though the table contains information about both types, we are actually presenting two independent evaluations, one for men and one for women. This implies that making comparisons by rows is meaningless, except for the coefficient of variation that shows that there is larger diversity between women than between men. We observe that life satisfaction declines with age, except for the older group of men. We also find here a much larger variability in the balanced worth than in average values for both types.

Table 8 provides the separate evaluation by groups. In spite of having a single table, we are actually presenting four separate evaluations. In this case the only meaningful comparisons are by rows (between men and women for each particular age group). Women are more satisfied with life than men for all age groups except the oldest one (this partly reflects the differences in life expectancy). We also find here that the balanced worth discriminates more than average values: the relative differences between men and women by age groups, according to the balanced worth, are 4% for the first group, 11% for the second, 2% for the third and – 10% for the last one. The corresponding values for the means are 1, 3, 0 and – 4%.

Finally, we present the results of the joint evaluation. Now each of the cells defined by age and gender is considered as a group and evaluated accordingly. Consequently, we can compare young women with old men, young women with old women, or young men with old men, for example. Table 9 provides the results. Note that the inclusion of all those population subgroups changes the values of the separate evaluation by types and age groups. Yet all qualitative traits are maintained: women fare better than men in all age groups except the older one, life satisfaction declines with age except for the older men, and the balanced worth presents a much larger variability than the mean values (about four times).

4 Final Remarks

The balanced worth provides an index that evaluates the relative goodness of a series of outcome distributions in terms of the likelihood of getting better or equal results. The key value judgement is that of comparing pairs of groups in terms of the probability that a random extraction from one of them yields a better or equal outcome than one random extraction from the other. The balanced worth corresponds to a consistent application of this notion for any number of groups.

There are several aspects of this evaluation method that deserve to be discussed in order to better understand its nature and applicability.

4.1 The Balanced Worth and the Worth

The balanced worth can be regarded as a modification of the concept of *worth*, introduced in Herrero and Villar (2013) and applied subsequently in a series of empirical problems, as already mentioned. The worth is defined as the consistent extension of the binary principle that evaluates the relative performance of two groups proportionally to their corresponding domination probabilities. That is, $v_1/v_2 = p_{12}/p_{21}$. This extension yields the following evaluation for each group in the general case:

$$v_i = \frac{\sum_{j \neq i} p_{ij} v_j}{\sum_{j \neq i} p_{ji}}, \quad i = 1, 2, \dots, g \quad (10)$$

The obvious difference between Eqs. (2) and (10) is that the second does not include the probability of ties in the evaluation. This makes the evaluation concentrate on the part of the distribution in which the groups differ and ignore that in which they are similar. This implies that the worth may strongly overestimate those differences when e_{ij} is large [let us remind here that $p_{ij} + p_{ji} + e_{ij} = 1$, so that Eq. (10) does not distribute all the probability mass between the groups whereas Eq. (2) does so].

4.2 Categories

The balanced worth requires very little information: the matrix of relative frequencies. This is why it can be naturally applied to evaluation problems involving categorical data, as the distribution of the elements of the population into the different categories is all we need. It therefore follows that the definition of those categories (how many and how inclusive they are) is key to obtain a sensible evaluation. Changes in the definition of those categories affect the matrix of relative frequencies and hence the final result. As all the elements within a category are indistinguishable, the more generic the category is, the less attention we pay to individual differences. And vice versa. A sensitivity analysis with some alternative specifications is advised when categories are part of the modelling choices. The easy computation of the balanced worth makes of this an immediate exercise.

4.3 Numerical Variables

Nothing prevents the application of this method to address problems involving numerical variables, either discrete or continuous. The empirical illustration on life satisfaction in Spain is a good example of that possibility. Yet one has to be careful when dealing with

numerical variables because they are to be interpreted as indexing attributes rather than as genuinely quantitative values. In particular, one has to bear in mind that this evaluation procedure does not compute the differences in the magnitude of the achievements, but just their distribution between the ordered categories.⁵

4.4 Relative Evaluation

The balanced worth is an index that provides a relative evaluation of the performance of a collection of groups, which means that the value attached to each group depends on all the groups with which it is compared. At the limit, the balanced worth is not defined when there is a single group involved. This property implies that the set of groups being compared should have something in common that makes analysing their relative behaviour relevant; otherwise the evaluation will be formally correct but of no interest. Deciding the groups that enter the comparison, therefore, matters. This is rather natural in some problems (e.g. the regions of a country), whereas it is a modelling choice in others. Be it as it may, the number and nature of the groups involved could affect the evaluation of each participant. That is, the relative evaluation of any two groups may be altered by considering or not a third party.

4.5 Endogenous Weighting

It might be tempting to think of the balanced worth as an endogenous way of attaching weights to the different categories or levels of performance, so that the result is a sort of weighted average. This is not (and cannot be) the case. The balanced worth cannot be identified with any method that attaches weights to the levels in the distributions being compared. Our evaluation criterion takes a different venue that cannot be formulated in terms of weights.

4.6 Further Research

The applications of the worth referred to in Sect. 1 serve also to illustrate the types of problems that can be addressed with the balanced worth. There are many other fields in which this instrument may be useful. We shall refer to some additional developments on which the authors are presently working.

4.6.1 *Evaluating the Quality of the Labour Market*

A clear way of enriching the analysis of the labour market is by getting an estimate of the overall evolution of the labour force in different categories of employment and unemployment. In particular, we are analysing the impact of the economic crisis in the Spanish

⁵ In the empirical application regarding life satisfaction individual answers have been grouped into a rougher set of categories. One may reasonably wonder what the purpose is of losing information by grouping those data into broader categories when we have all the individual numerical responses. The main reason is that when dealing with subjective evaluations in terms of numerical scales, there is no guarantee that numbers mean the same for different people (your 7 and my 7 may well represent very different things). Moreover, individual scales need not be linear (i.e. an evaluation 8 need not be twice one of 4, even for a single individual). Grouping numerical answers into categories may thus help illuminate some structural features of the groups, enhance robustness and reduce the comparability assumptions required.

labour market taking into account the distribution of the labour force in the following categories: permanent contracts, temporal contracts, workers unemployed less than three months, workers unemployed between 3 and 12 months, workers unemployed between 1 and 2 years, and workers unemployed for more than 2 years. Preliminary results show that the economic recovery in Spain is less brilliant than the figures of unemployment suggest, due to the worsening of the employment conditions and the persistence of long run unemployment.

4.6.2 Comparing Assets with Unknown Returns

Suppose we have to compare a series of assets whose future returns are not known with precision. Our information refers to the likelihood of the different states of the world and whether the returns will be higher or lower in one state than in the other. Suppose, for the sake of illustration, that we consider 5 possible states of the world, arranged from best to worse. We can identify each asset (that here plays the role of a group) with the corresponding distribution of probabilities over those states of the world (that play the role of categories). We can use then the balanced worth to rank the different assets and make our investment decision.

4.6.3 Evaluating Pain Relief in Clinical Trials

Most clinical trials regarding the effectiveness of pain relief treatments are based on the patients' responses to a 0–10 pain scale. Those numbers are aggregated and treated statistically as numerical variables to assess the performance of alternative drugs or treatments. Yet those scales are intrinsically qualitative and hardly comparable interpersonally. Treating the grades of the pain scale as categories and analysing the distribution of the patients within those categories with the balanced worth provides a sounder evaluation of the outcomes.

4.6.4 Multidimensional Evaluation

On a different venue, there is the case in which the populations are to be evaluated with respect to more than one variable. The extension of this method to the multidimensional case is one of the lines of research we are exploring.

Acknowledgements Thanks are also due to Héctor García Peris, for his help in developing the algorithm that computes the evaluation, and to an anonymous referee for very helpful comments and suggestions. Funding was provided by the Spanish *Ministerio de Economía y Competitividad* (Grant Nos. ECO2015-65408-R, ECO2015-65820-P).

Appendix: Existence and uniqueness of the balanced worth

Here we prove that the balanced worth always exists and that, under very general conditions, it is unique and strictly positive.

Definition Given a problem **A**, a group j is *fully dominated* in **A**, if for all $i \neq j$ it happens that $p_{ij} = 1$.

When a group j is fully dominated, $p_{ji} = e_{ji} = 0, \forall i \neq j$. That is, all individuals in group j belong to a lower level than any other individual in the rest of the groups. Note that, in practice, the probability of finding a fully dominated group is zero.

Let us formally state that a solution to the system of g equations with g unknowns (3), always exists.

Theorem 1 *Let A be an evaluation problem. Then:*

- (i) *A vector $\mathbf{w} \in \mathbb{R}_+^g$ exists that solves equation system (3). That is, a vector \mathbf{w}^* such that:*

$$w_i^* = \frac{\sum_{j \neq i} \left(p_{ij} + \frac{e_{ij}}{2} \right) w_j^*}{\sum_{j \neq i} \left(p_{ji} + \frac{e_{ji}}{2} \right)}, \quad i = 1, 2, \dots, g$$

- (ii) *If no group is fully dominated, then the solution is unique (up to a scalar multiplication) and strictly positive.*

Proof

- (i) Let $W = \{ \mathbf{w} \in \mathbb{R}_+^g / \sum_{i=1}^g w_i = g \}$ and consider the function $\varphi : W \rightarrow \mathbb{R}$, given by:

$$\varphi_i(\mathbf{w}) = w_i - \frac{1}{g-1} \left(w_i \sum_{j \neq i} \left(p_{ji} + \frac{e_{ji}}{2} \right) - \sum_{j \neq i} \left(p_{ij} + \frac{e_{ij}}{2} \right) w_j \right)$$

As $\sum_{j \neq i} \left(p_{ji} + \frac{e_{ji}}{2} \right) \leq g - 1$, we have:

$$\varphi_i(\mathbf{w}) \geq w_i - w_i + \frac{1}{g-1} \sum_{j \neq i} \left(p_{ij} + \frac{e_{ij}}{2} \right) w_j \geq 0$$

Moreover,

$$\sum_{i=1}^g \varphi_i(\mathbf{w}) = g - \frac{1}{g-1} \left(\sum_{i=1}^g w_i \sum_{j \neq i} \left(p_{ji} + \frac{e_{ji}}{2} \right) - \sum_{i=1}^g \sum_{j \neq i} \left(p_{ij} + \frac{e_{ij}}{2} \right) w_j \right)$$

Note that, by construction,

$$\sum_{i=1}^g w_i \sum_{j \neq i} \left(p_{ji} + \frac{e_{ji}}{2} \right) = \sum_{i=1}^g \sum_{j \neq i} \left(p_{ij} + \frac{e_{ij}}{2} \right) w_j$$

which means that $\sum_{i=1}^g \varphi_i(\mathbf{w}) = g$ and hence that function φ maps W into itself. As it is a continuous function and W is a compact convex set, Brouwer’s Theorem (e.g. Border 1989), ensures the existence of a fixpoint, $\mathbf{w}^* = \varphi(\mathbf{w}^*)$. That is,

$$w_i^* = w_i^* - \frac{1}{g-1} \left(w_i^* \sum_{j \neq i} \left(p_{ji} + \frac{e_{ji}}{2} \right) - \sum_{j \neq i} \left(p_{ij} + \frac{e_{ij}}{2} \right) w_j^* \right)$$

and, therefore,

$$w_i^* = \frac{\sum_{j \neq i} \left(p_{ij} + \frac{e_{ij}}{2} \right) w_j^*}{\sum_{j \neq i} \left(p_{ji} + \frac{e_{ji}}{2} \right)}, \quad i = 1, 2, \dots, g$$

(ii) Assume now that there is no fully dominated group, that is, $(p_{ij} + \frac{e_{ij}}{2}) > 0 \forall i, j$. Then, the solutions must be strictly positive. This is so because both numerator and denominator are strictly positive. To prove uniqueness, suppose there are two strictly positive vectors, w, y , that solve the equation system (3). Then, we can write:

$$\sum_{j \neq i} \left(p_{ji} + \frac{e_{ji}}{2} \right) = \frac{\sum_{j \neq i} \left(p_{ij} + \frac{e_{ij}}{2} \right) w_j}{w_i} = \frac{\sum_{j \neq i} \left(p_{ij} + \frac{e_{ij}}{2} \right) y_j}{y_i}, \quad i = 1, 2, \dots, g$$

For a given i , this expression can be rewritten as:

$$A = \sum_{i=1}^{g-1} B_i x_i = \sum_{i=1}^{g-1} B_i z_i$$

where all terms are strictly positive, with $x_j = w_j/w_i, z_j = y_j/y_i$. But this is the equation of a hyperplane with a given normal, which means that vectors x and z are to be proportional. That is, the solution is unique up to the choice of units. □

References

Albarrán, P., Herrero, C., Ruiz-Castillo, J., & Villar, A. (2017). The Herrero-Villar approach to citation impact. *Journal of Informetrics*, 11(2), 625–640.

Bellù, L. G., & Liberati, P. (2005). *Social welfare analysis of income distributions ranking income distributions with crossing generalised Lorenz curves*. Rome: Food and Agriculture Organization of the United Nations.

Border, K. C. (1989). *Fixed point theorems with applications to economics and game theory*. Cambridge: Cambridge University Press.

Bourguignon, F., Ferreira, F. H. G., & Leite, P. G. (2008). Beyond Oaxaca–Blinder: Accounting for differences in household income distributions. *Journal of Economic Inequality*, 6(2), 117–148.

Chakravarty, S. R., & Silber, J. (2007). A generalized index of employment segregation. *Mathematical Social Sciences*, 53(2), 185–195.

Crespo, J. A., Li, Y., & Ruiz-Castillo, J. (2013). The measurement of the effect on citation inequality of differences in citation practices across scientific fields. *PLoS ONE*, 8(3), e58727. <https://doi.org/10.1371/journal.pone.0058727>.

Cuhadaroglu, T. (2013). *My group beats your group: Evaluating non-income inequalities*. St Andrews: W.P. School of Economics and Finances, University of St Andrews.

Echenique, F., & Fryer, R. G. (2005). *On the measurement of segregation* (Vol. w11258). Cambridge: National Bureau of Economic Research.

Eurostat. (2015). *Quality of life. Facts and views*. Luxemburg: Publication Office of the European Union.

Frankel, D. M., & Volij, O. (2011). Measuring school segregation. *Journal of Economic Theory*, 146(1), 1–38.

Gallén, M. L., & Peraita, C. (2015). A comparison of corporate social responsibility engagement in the OECD countries with categorical data. *Applied Economics Letters*, 22(12), 1005–1009.

Gonzalez-Diaz, J., Hendrichx, R., & Lohmann, E. (2014). Paired comparison analysis: An axiomatic approach to ranking methods. *Social Choice and Welfare*, 42(1), 139–169.

Grannis, R. (2002). Segregation indices and their functional inputs. *Sociological Methodology*, 32(1), 69–84.

- Herrero, C., Méndez, I., & Villar, A. (2014). Analysis of group performance with categorical data when agents are heterogeneous: The evaluation of scholastic performance in the OECD through PISA. *Economics of Education Review*, *40*, 140–151.
- Herrero, C., & Villar, A. (2013). On the comparison of group performance with categorical data. *PLoS ONE*, *8*(12), e84784. <https://doi.org/10.1371/journal.pone.0084784>.
- Laband, D. N., & Piette, M. J. (1994). The relative impacts of economics journals: 1970–1990. *Journal of Economic Literature*, *32*(2), 640–666.
- Laslier, J. (1997). *Tournament solutions and majority voting*. Berlin: Springer.
- Li, F., Yi, K., & Jests, J. (2009). *Ranking distributed probabilistic data*. SIGMOD'09, June 29–July 2.
- Liebertson, S. (1976). Rank-sum comparisons between groups. *Sociological Methodology*, *7*, 276–291. <https://doi.org/10.2307/270713>.
- Martínez-Mekler, G., Martínez, R. A., del Río, M. B., Mansilla, R., Miramontes, P., et al. (2009). Universality of rank-ordering distributions in the arts and sciences. *PLoS ONE*, *4*(3), e4791. <https://doi.org/10.1371/journal.pone.0004791>.
- Palacios-Huerta, I., & Volij, O. (2004). The measurement of intellectual influence. *Econometrica*, *72*(3), 963–977.
- Pavot, W., & Diener, E. (2008). The satisfaction with life scale and the emerging construct of life satisfaction. *The Journal of Positive Psychology*, *3*, 137–152.
- Pinski, G., & Narin, F. (1976). Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. *Information Processing and Management*, *12*(5), 297–312.
- Reardon, S. F., & Firebaugh, G. (2002). Measures of multi-group segregation. *Sociological Methodology*, *32*, 33–76.
- Rosvall, M., & Bergstrom, C. T. (2007). An information-theoretic framework for resolving community structure in complex networks. *Proceedings of the National Academy of Sciences*. <https://doi.org/10.1073/pnas.0611034104>.
- Sheriff, G., & Maguire, K. (2013). *Ranking distributions of environmental outcomes across population groups*. National Center for Environmental Economics. August.
- Shorrocks, A. F. (1983). Ranking income distributions. *Economica*, *50*(197), 3–17.
- Slutzki, G., & Volij, O. (2006). Scoring of web pages and tournaments. *Social Choice and Welfare*, *26*(1), 75–92.
- Torregrosa, R. (2015). Medición y evolución del sentimiento autonómico en España. *Investigaciones Regionales- Journal of Regional Research*, *33*, 53–70.
- Veenhoven, R. (1991). Is happiness relative? *Social Indicators Research*, *24*, 1–34.
- Villar, A. (2014). Education and cognitive skills in the Spanish adult population. Intergenerational comparison of mathematical knowledge from PIAAC data. *Advances in Social Sciences Research Journal*, *1*(1), 72–88.
- Yalonetzky, G. (2012). A dissimilarity index of multidimensional inequality of opportunity. *The Journal of Economic Inequality*, *10*(3), 343–373.