PERGAMON

Omega 30 (2002) 227–242

omega
The International Journal of Management Science

www.elsevier.com/locate/dsw

Facilitating bid evaluation in public call for tenders: a socio-technical approach

Carlos A. Bana e Costa[a,b,*], Émerson C. Corrêa[c], Jean-Marie De Corte[d], Jean-Claude Vansnick[d]

[a] *Centre of Management Studies (CEG-IST), Instituto Superior Técnico, Av. Rovisco Pais, 1049-001 Lisbon, Portugal*
[b] *Operational Research Department, London School of Economics, Houghton Street, London WC2A 2AE, UK*
[c] *Olympus Consulting S.L., C/Torrelaguna, 67 3° A, 28027 Madrid, Spain*
[d] *Centre de Recherche Warocqué, Université de Mons-Hainaut, Place du Parc, 20, 7000 Mons, Belgium*

Received 3 November 2000; accepted 2 March 2002

Abstract

A specific multicriteria socio-technical approach to facilitating bid evaluation processes is presented and several issues that warrant its use are discussed. Some real-world interventions in international public call for tenders illustrate practical aspects of structuring criteria and creating a computer-based additive value model in direct interaction with Evaluation Committees responsible for bid evaluation, supported by the MACBETH approach. © 2002 Elsevier Science Ltd. All rights reserved.

Keywords: Public call for tenders; Multicriteria bid evaluation; MACBETH; Real-world cases

1. Introduction

Several specific characteristics of bid evaluation processes in public call for tenders warrant the use of a methodology to back up decision-making. For instance, it is within the discretion of the Evaluation Committee (typically a decision-making group of 5–7 experts and other key-players) to decide, though not in an arbitrary manner. This makes it essential for them to be duly qualified to provide a clear, unambiguous explanation, to ensure that their decisions comply with the principles safeguarding equal treatment for all bidders and transparency in the allocation of public resources.

We believe that a multicriteria decision aid (MCDA) is the most suited approach for this type of public decision context. This is based on our firm belief that the explicit introduction of various criteria is a better method of making a good decision when confronted by a multidimensional problem than merely considering one single evaluation criterion such as the lowest price criterion—see [1] for an extensive survey on the use of MCDA techniques for competitive bidding. However, facilitating bid evaluation cannot be reduced to a technical process of weighting criteria, rating options, and then simply applying an aggregation procedure to obtain a final ranking of the bids. The Evaluation Committee's work has a fundamental social component. Therefore, in conjunction with providing the appropriate technical support, the facilitator should help a decision-making group composed of people from diverse professional backgrounds and value systems, and, consequently, with different perspectives about the value of the options under analysis (often meeting each other for the first time) to discuss their viewpoints in order to develop a shared understanding of the key issues and a sense of common purpose, and exercise, together, their expert judgements in order to come to an agreement on the best bid(s). It is our experience that Decision Conferencing (cf. [2]), a socio-technical process composed of a series of

* Corresponding author. Operational Research Department, London School of Economics, Houghton Street, London WC2A 2AE, UK. Tel.: +44-207-9556-193; fax: +44-209-7556-885.

E-mail addresses: c.bana@lse.ac.uk (C.A. Bana e Costa), ecorrea@uhcmadrid.org (E.C. Corrêa), DeCorte@umh.ac.ne (J.-M. De Corte), Vansnick@umh.ac.be (J.-C. Vansnick).

PII: S0305-0483(02)00029-4

"decision conferences" (that is, intensive "face-to-face" facilitated group-working sessions lasting each one from 1 to 2 days, intermediate with "off-line" data gathering and processing) is the adequate framework for the development of a multicriteria bid evaluation.

However, there are some methodological and technical restrictions in using a multicriteria approach in public call for tenders. One, for instance, is the fact that the awarding authority is legally obliged to publish, in the announcement of the call for tenders, the evaluation criteria and their respective weights, or at least their order of relative "importance".

As a result, once the criteria have been published, it is legally unacceptable (according to some legal opinions, with which we agree) to proceed to any evaluation of the bids according to a given criterion, based on the application of a weighted sum of scores assigned to them in aspects that were not indicated in the call for tenders as sub-criteria. This restriction makes the evaluation of bids by criteria such as *Technical quality* difficult, particularly in projects or equipment of great complexity, since there are usually numerous indicators and features that should be taken into consideration when evaluating based on such a criterion.

Moreover, the requirement to publish criteria (and sub-criteria) weights implies that they must be defined before the bids are known. This introduces another technical problem, since the elicitation of "weights" without reference to the domains of variation of the impacts on criteria is theoretically incorrect and has no mathematical meaning in the framework of an additive aggregation model.

Faced with these and other problems to be discussed later, we have developed a specific multicriteria methodology that has been applied (validated and improved) in numerous bid evaluation processes, in which some of us have acted as decision analysts and facilitators. It is upon this experience of intensive interaction with Evaluation Committees that the present paper is based. The proposed methodology includes two main phases: structuring and evaluation, addressed in Sections 2 and 3, respectively. The evaluation phase encompasses the use of scientifically correct yet simple techniques, namely the MACBETH approach (*M*easuring *A*ttractiveness by a *C*ategorical *B*ased *E*valuation *T*echnique)—cf. [3–5]. Specific multicriteria software, namely M-MACBETH, is used to support the "on-the-spot" creation of a computer-based additive value model and to perform sensitivity and robustness analyses on the results derived from the model's application. These analyses are essential for drawing up recommendations in regard to the (relative or intrinsic) attractiveness of bids; indeed, the evaluation procedure should never be concluded until the sensitivity/robustness of the recommendations coming out of the model's application have been discussed. This is essential to guarantee that the model is a "requisite model" (in the sense of Phillips [6]).

As detailed in the next sections, our approach to support bid evaluation can be viewed as a package of activities (Fig. 1) to be developed in decision conferences. The activities described below should be understood as generic and, therefore, to be adapted to each particular type of call for tenders in its respective *Evaluation Regulations*:

(a) *Characterisation of the decision context*, by identifying the actors involved and process constraints (see example-case description in Section 4.1);
(b) *Definition of screening and evaluation criteria*, through the identification and structuring of all aspects considered relevant to the analysis and evaluation of bid submissions (Section 2.1);
(c) *Construction of a descriptor of impacts for each evaluation criterion*, based on the indicators and the characteristics that allow these criteria to be made operational (Section 2.2);
(d) *Determination of relative weights* that make operational the notion of "relative importance" of the evaluation criteria in the framework of an additive aggregation model, applying MACBETH (Section 3);
(e) *Impact appraisal and partial evaluation of the bids* for each criterion, applying MACBETH (Section 4.1);
(f) *Calculation of the overall value of each bid*, through the additive aggregation model (Section 4.1):
(g) *Sensitivity and robustness analyses of the results* in order to allow an appropriate drafting of recommendations (Section 4.2).

2. Structuring phase

2.1. Definition of screening and evaluation criteria

A criterion is a tool used to evaluate bids in terms of a certain point of view or concern considered as fundamental or key by the decision-making group. There are two types of criteria: screening and evaluation criteria. Screening criteria represent the deliberate intention to make bidders comply with thresholds of admissibility and to only proceed to comparative evaluation of bids from bidders who do so. Each evaluation criterion is (should be) an independent axis of comparative evaluation. This is the reason why, very often, several interrelated concerns have to be grouped in the same criterion (examples related with ordinal and cardinal independency requirements are given in Sections 2.2.1 and 3, respectively).

The set of evaluation criteria should be as concise as possible. Unfortunately, it is not uncommon to come across a call for tenders in which different indicators are defined as distinct criteria when they in fact represent the same concern. This introduces redundancy to the model, the consequence of which will be the overvaluation of what should be only one criterion.

Several levels of specification may be considered in defining the evaluation criteria; in these cases, it is useful to represent the criteria and sub-criteria in a tree structure, as displayed in Fig. 2. The value tree should be complete,

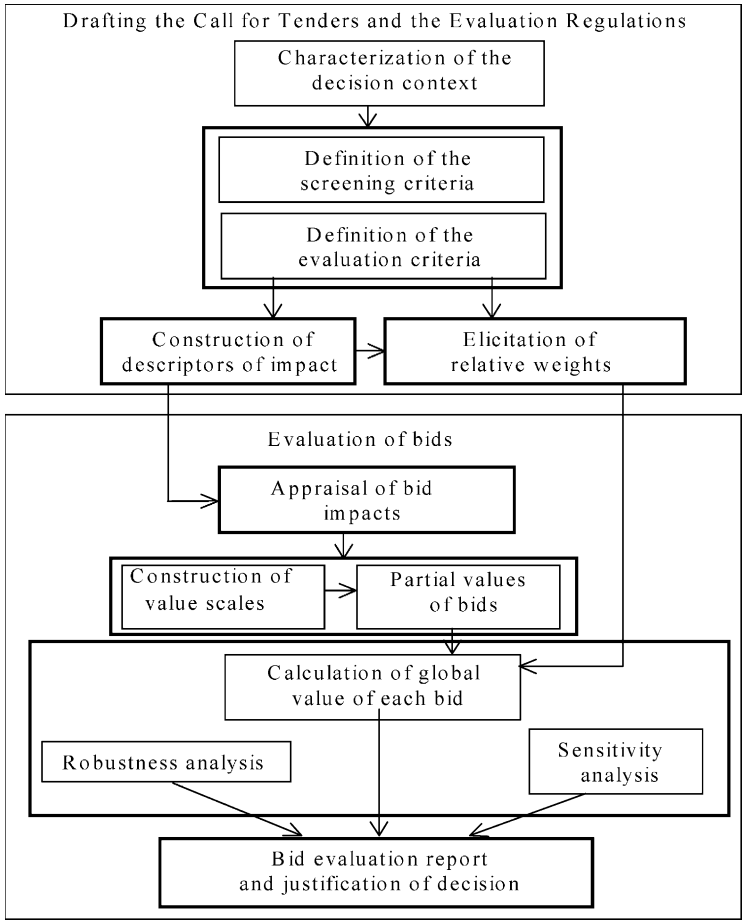


Fig. 1. Methodological diagram. Each bold-framed box typically corresponds to one (or more, depending on the complexity of the case) decision conference.

capturing all of the fundamental aspects for evaluating the bids. For instance, in the case of Fig. 2c, the initial set of criteria did not include the criterion *Credibility of cost*, which reflects a risk concern. Indeed, when the Evaluation Committee was faced with the question "if two bids were indifferent in each of the criteria already defined, is there any reason to consider one globally better than the other?", the answer was "yes, if we knew that the estimated costs were more credible in one of them". Notice that the fact that this criterion was difficult to operationalise does not justify its exclusion, as this would lead to an incomplete evaluation model.

In cases of great complexity, numerous primary evaluation aspects are present, such as the technical requirements mentioned in the Terms of Reference of the call for tenders or other aspects that emerged in the initial decision conference devoted to structuring. How detailed should the tree of evaluation criteria be? The search for an adequate answer can be facilitated by adopting processes that help to identify clusters of linked aspects, such as "post-it sessions" (see [7, Section 12.4.2]) and "cognitive mapping" [8,9]. We use to represent the results of the processes in a "table of concerns" similar to Table 1. For each criterion or sub-criterion, we distinguish two possible hierarchical levels of specification of primary aspects: "indicators" and, when appropriate, "characteristics". Table 1 reproduces the section related to the *Work methodology* criterion in Fig. 2b. For instance, the sub-criterion *Technical procedures* has six indicators each one composed of several characteristics, whereas only one level of specification was necessary for the two other sub-criteria.

The essential activity of defining criteria has an added importance in public calls for tenders, since, as stated in Section 1, the criteria cannot be changed after the publication of the call for tenders and the use of weighted sums at the level of the indicators or characteristics is not permissible by law, as they were not announced as criteria or sub-criteria. Let us examine in Section 2.2.2 how to overcome this problem.

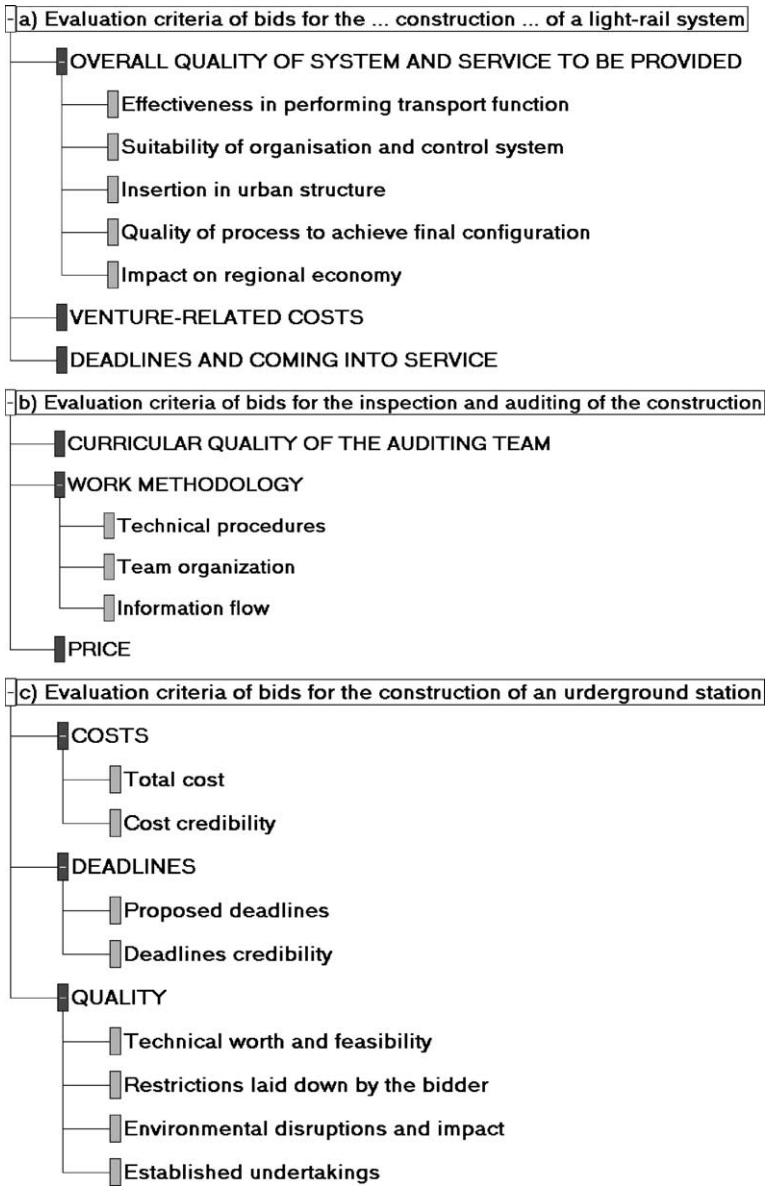


Fig. 2. Evaluation criteria of the International Public Calls for Tenders for: (a) the "design, construction, equipment providing, financing and short-term operation, of a light-rail system in the Metropolitan Area of Porto"; (b) "inspection and auditing of the construction of the Porto light-rail system"; (c) "the design and construction of the Terreiro do Paço station" (a and b issued by Metro do Porto S.A. and c by Metropolitano de Lisboa, E.P.).

2.2. *Operationalisation of evaluation criteria*

2.2.1. *Definition of descriptors of plausible impacts*

To make a criterion (or sub-criterion) operational for bid evaluation, a "descriptor" of impacts is associated with it. By definition, a descriptor is an ordered set of (quantitative or qualitative) plausible impact levels. Depending on the context, an evaluation criterion can be operationalised by a natural, proxy or indirect, or a constructed descriptor. If it is possible to define natural descriptors, this is the appropriate choice, since the more objective the descriptors are, the less ambiguous the criteria will be and, therefore, the less controversial the evaluation model. This facilitates the justification of the final decision and meets the requirement of no arbitrariness. However, the existence of a natural descriptor does not imply that it is necessarily the best one to operationalise the respective criterion. For example, the *Cost of acquisition* of certain equipment may be associated with the

Table 1
Example of a table of concerns (Metro do Porto S.A.)

Criteria	Sub-criteria	Indicators	Characteristics
Work methodology	Technical procedures	Project revision	Methodology
		Information management	System reliability
			Information back-up
			User friendship
			Root support for the MP document encryption scheme
			Location and access to archives
		Deadline management	Procedures for approval of work plans
			Databases
			Resource assignment control system
			Probability evaluation of deadline fulfilment
			Method followed in correcting deviations from plans
			Software used in deadline management
		Quality management	State of implementation of quality system in companies
			Appropriateness of proposed quality system
			Appropriateness of quality system implementation schedule
		Safety management	State of implementation of work safety system in firm
			Appropriateness of proposed work safety system
			Appropriateness of work safety system implementation schedule
		Costs control	Work costs database
			Cost prevision system
			Methodology to verify bills and quantities
			Methodology to establish new prices
			Price revision
			Cost-term interconnection
	Team organisation	Organisation chart	
		Team size	
		Diagram of personnel workload	
	Information flow	Process of information collection	
		Flowchart of information circuits	

(natural) descriptor *proposed price* if the payment is made all at once, or the (indirect) descriptor *present value of the proposed price* if the payment is to be made in instalments, or even *revised present value of the proposed price* in contexts of high inflation. In a call for tenders in which the quantities to be provided or carried out may vary between bids, or in a call for tenders "by price series", the descriptor may be more sophisticated. This was the case of the call for tenders for "inspection and auditing of the construction of the light-rail system" (Fig. 2b), in which a "Typical-Team" was defined in the Call for Tenders and the descriptor used for bid comparisons was (*comparison*) *price resulting from the application of proposed unit prices to the Typical-Team*.

On the other hand, constructed descriptors combining several indicators can be used to avoid problems of ordinal dependence. For example, in the call for tenders put out by the Metropolitano de Lisboa, E.P. to carry out a work order integrated in the construction of the Alameda-Expo line in Lisbon, the *Deadlines* criterion was defined with two indicators, *Completion deadline* and *Viaduct completion deadline* (the latter a prerequisite to start other work contracts). Why was each of these indicators not taken as one distinct sub-criterion? Because the conclusion of the work contract is not "isolable" from the conclusion of the viaduct, and the final deadline may depend on the intermediate deadlines due to the sequential relationships between the activities succeeding or preceding the construction of the viaduct. The two deadlines were, thus, not independent and had to be assessed together. This is why the descriptor associated with the criterion *Deadlines* was a constructed descriptor, whose impact levels were deadline profiles corresponding to two indicators, profiles which were holistically rank-ordered by the Evaluation Committee.

2.2.2. Definition of levels of reference "Good" and "Neutral"

There are three reasons for recommending the identification of two reference levels ("good" and "neutral") of intrinsic value in each criterion, that operationalise the idea of a good bid and a neutral bid (that is, neither attractive nor repulsive):

1. Experience has revealed that the effort required to identify good and neutral levels contributes significantly to

Table 2
Example of intrinsic reference levels of price (Metro do Porto S.A.)

Level	Description
Good	A bid with a price that is 300 million Portuguese escudos lower than the average (comparison) prices among the bids presented in the call for tenders
Neutral	A bid with a price that is 100 million Portuguese escudos higher than the average (comparison) prices among the bids presented in the call for tenders

Table 3
Example of reference levels for indicators (Metro do Porto S.A.)

Indicator	Good	Neutral
Organisation chart	Well-defined areas of responsibility and work assignment that tend to minimise zones of divided responsibility. Hierarchical dependencies clearly assigned and not shared. Existence of automatic responsibility delegation in case of temporary absences	Well-defined areas of responsibility. Some shared hierarchical dependencies, but clearly justified
Team size	Team sizing that ensures efficient performance of auditing tasks, allowing a reduction in the size of the Typical-Team equivalent to 10%	Team sizing equivalent to the Typical-Team and that ensures efficient performance of auditing tasks
Diagram of personnel workload	Totally coherent with the physical planning of the work	Diagram with no significant incoherence

the intelligibility of the criteria. It is one thing to say that a bid submission is better than another in the *Price* criterion, for example, yet, it is quite another to specify what is meant by a good price or a neutral price. Table 2 shows how these levels were defined in the Evaluation Regulation for the *Price* criterion in Fig. 2b.

2. An explicit statement regarding good and neutral levels of reference makes it possible to objectify the notion of intrinsic attractiveness of each bid, assigning it to one of the following categories:
   - *very positive bid*, when it is at least as attractive as a fictitious good bid;
   - *positive bid*, if it is at least as attractive as a neutral fictitious bid, but less attractive than a fictitious good bid;
   - *negative bid*, if it is less attractive than a neutral fictitious bid.

   Making the reference levels explicit, rather than simply producing a relative attractiveness ranking through a comparative evaluation of the bids, permits the determination of the intrinsic value of each of them, which helps to avoid situations in which an inappropriate bid is chosen, even if it is the best bid present in a call for tenders (in this case, the best of a set of "bad" bids).
3. Defining the two reference levels allows the use of a criteria-weighting procedure that simultaneously matches specific characteristics of public calls for tenders and is valid in the theoretical framework of the application of an additive aggregation model, as we will see in Section 3.

When a criterion (or sub-criterion) integrates several "indicators", a neutral (respectively, good) level can have several definitions, for instance: "all indicators at their respective neutral levels (respectively, good)". Note that it is also necessary to make explicit the reference levels for the indicators. Table 3 provides an example for the indicators of the sub-criterion *Team organisation* described in Table 1, as stated in the Evaluation Regulations.

The degree of complexity involved would increase if the indicators were sets of "characteristics". For this reason, we developed a procedure, called the "determinants technique" which could be applied by the Evaluation Committee. The methodological basis of this technique consists of a sequence of key procedure rules, like:

1. Establish two reference levels, "satisfactory (+)" and "neutral (o)", in each characteristic;
2. Classify each characteristic as "determinant" (D), "important" (I) or "secondary" (S). A characteristic will be "determinant" if a bid being negative (worse than neutral) in that characteristic is a necessary and sufficient condition for the bid to be considered negative (worse than neutral) in the respective indicator. Note that this idea is in line with the notion of "veto power" used to model non-compensatory situations (cf. [10]);

Table 4
Example of application of the *determinants technique* (Metro do Porto S.A.)

Characteristic (type)	(+) Satisfactory	(o) Neutral
Cost prevision system (D)	The costs prevision system takes into account the physical planning of the works	The costs prevision system is only based on progress verified in the financial time-chart
Cost-term inter-connection (D)	There is only one database for costs and deadlines	There is validation between the two management systems
Work costs database (I)	Database that enables price estimates through a composed price methodology	Database adapted to the progress of the works
Methodology to verify bills and quantities (I)	Totally automated methodology	Methodology based on totally manual processes
Methodology to establish new prices (I)	Market prices are systematically collected and an analysis is also made by composed price	Market prices are systematically collected
Price revisions (S)	Price revisions are verified automatically and there is a system to predict price indexes	Price revisions are verified automatically

3. Define the "good" level in the indicator by: "All determinant characteristics satisfactory and a majority of important characteristics satisfactory"; define the "neutral" level in the indicator by: "A majority of determinant and important characteristics neutral, without any characteristic negative".

Table 4 provides an example of this situation for the characteristics of the indicator *Costs control* (from the sub-criterion *Technical procedures*) described in Table 1, as stated in the Evaluation Regulations.

3. Weighting the evaluation criteria

3.1. Discussion

Following our methodology for bid evaluation, an additive value model of the type:

$$V(b) = \sum_{j=1}^{n} k_j v_j(b) \quad \text{with} \sum_{j=1}^{n} k_j = 1$$

$$\text{and } k_j > 0 \text{ and } \begin{cases} v_j(\text{good}_j) = 100, \\ v_j(\text{neutral}_j) = 0 \end{cases}$$

will be constructed to aggregate the partial values $v_j(b)$ of each bid b in the criteria ($j=1,\ldots,n$) and calculate its overall value $V(b)$ simultaneously taking into account the n criteria. If sub-criteria are present, the procedure is applied firstly for each group of sub-criteria sharing the same parent criterion. The parameters k_j are the scaling factors—commonly known as "weighting coefficients" or relative "weights"—that allow partial value units to be transformed into overall value units, through some form of operationalisation of the notion of trade-off: how much the Evaluation Committee considers necessary to improve impact in one criterion to compensate a decrease of impact on another criterion.

The compensatory additive aggregation procedure is, certainly, the simplest and the most frequently used multicriteria method, allowing not just the ordering of the bids in terms of their overall attractiveness, but also judging their relative difference of attractiveness, that is, *to what extent* one bid is better than the other. That implies, in mathematical terms, that the v_j are cardinal scales and, from an "axiomatic" point of view, that, besides the ordinal property of "isolability", the evaluation criteria (and sub-criteria) also have to respect the more demanding property of "independence in the sense of the differences of attractiveness"—also called "additive independence". In our constructive and more pragmatic approach, these axiomatic considerations are directly linked to the possibility of constructing scales of partial value. Consider, for example, the indicators *Team size* and *Diagram of the personnel workload*, from the sub-criterion *Team organisation*, and the respective reference levels defined in Table 3. It is possible to construct, for each of these two indicators, an ordinal scale of partial value because: (1) between two bids that present personnel workload diagrams with the same level of coherence, the best will always be the one that allows a greater reduction in the Typical-Team, regardless of whether the level of common coherence is high or low; and (2) between bids that allow an equal reduction in the Typical-Team, the best will always be the one that presents the most coherent personnel workload diagram, regardless of whether the common reduction in the Typical-Team is greater or smaller. This possibility can be put in relationship with the property of ordinal independence of the indicators *Team size* and *Diagram of the personnel workload.* However, the two indicators are cardinally dependent, which is made evident in our approach by the impossibility to construct a cardinal scale of partial value on each of them. Indeed, the greater the coherence of the personnel workload diagram the more attractive a reduction of 10% in the Typical-Team. Consequently, when using the

additive aggregation model in a cardinal perspective, the two indicators cannot be considered as evaluation criteria nor weighted separately.

As "weights" are substitution rates, their determination will have to be made with reference to criteria impact scales. Otherwise, the weights are arbitrary and make no sense in the additive framework, as when determined directly by reference to the psychological and intuitive notion of "importance". Unfortunately, there is whole panoply of more or less popular direct weighting processes that ignore these considerations, and are therefore theoretically incorrect—[11] refers to this as the "most common critical mistake."

This is the reason why correct weighting procedures—like the classic "tradeoff procedure" [12] or the more pragmatic "swing weighting" [13,14]—base their scaling factors (weights) assessment on actors' answers to questions that require from them a comparison of reference alternatives, traditionally defined on the base of the best (most attractive) and worst (least attractive) impact levels of the options on the criteria.

Let us present an example, adapted from Bana e Costa and Vansnick [3]. Imagine a bid evaluation situation involving only *Completion Deadline* and *Global Price*, the best and worse deadlines proposed being, respectively, 35 and 40 months, and the best and worst price proposed, respectively, 75 and 100 million euros. Let (35 months, 100 million €) be the impact profile of a fictitious bid x and (40 months, 75 million €) that of the fictitious bid y. Suppose that, when confronted with the comparison between x and y, the Evaluation Committee judged x more attractive than y, which means that the awarding authority would be prepared to pay 25 million €, from 75 to 100 million €, to reduce the deadline from 40 to 35 months. In this case, the additive model would take, correctly, a value for the weight of the deadline criterion which is greater than that for the cost criterion (as long as the partial values attributed to the best and worst impacts are, obviously, always equal in all criteria, for example, 100 and 0, respectively). Assume now that, some bids were eliminated for violating some of the screening criteria, and that the worst deadline among the remaining bids is now 39 months, instead of 40, while the best deadline is now 36 months, instead of 35, maintaining the best and worst prices. In these circumstances, the Evaluation Committee should compare profile x' (36 months, 100 million €) with profile y' (39 months, 75 million €). It is not unrealistic to imagine that y' would now be judged preferable to x'. The weight of the deadline criterion would, then, become, necessarily, smaller than the weight of the cost criterion! So what would the "most important" criterion be?

This phenomenon—known in the literature as the influence of scale ranges on criteria weights (cf. [15])—is due, in the above example, to the alterations in the deadline interval limits and clearly shows how incorrect it is to directly attribute values to weights without reference to impacts.

In recalling the legal requirement to publish the evaluation criteria and their respective weights (or at least, the order of "importance") in the call for tenders, that is, necessarily before the bids are known, we can conclude that it is impossible in these situations to define weights, or simply their order, with reference to the swings from the worst to the best impacts of the bids according to the different criteria, since these impacts are not yet known at the time of preparing those documents. However, mathematically, any non-indeterminate system of $n-1$ equations corresponding to $n-1$ judgements between alternatives, allows us to calculate the n weights, as soon as the reference alternatives are based on two distinct anchoring impacts in each criterion. In this case, to overcome the problem raised by the requirement to calculate a priori weights, it is sufficient that the weighting procedure be based on impact levels *of intrinsic value*, such as "good" and "neutral" levels.

To order the criteria weights, it is sufficient to ask the Evaluation Committee: "Consider a fictitious bid (N), with neutral impact level in all criteria. If it were possible to improve the impact from neutral to good in only one criterion, maintaining all of the others at their neutral level, for which criterion would this swing be most attractive? And the next most attractive?" Repeating this question until no criteria are left, leads directly to the order of the weights.

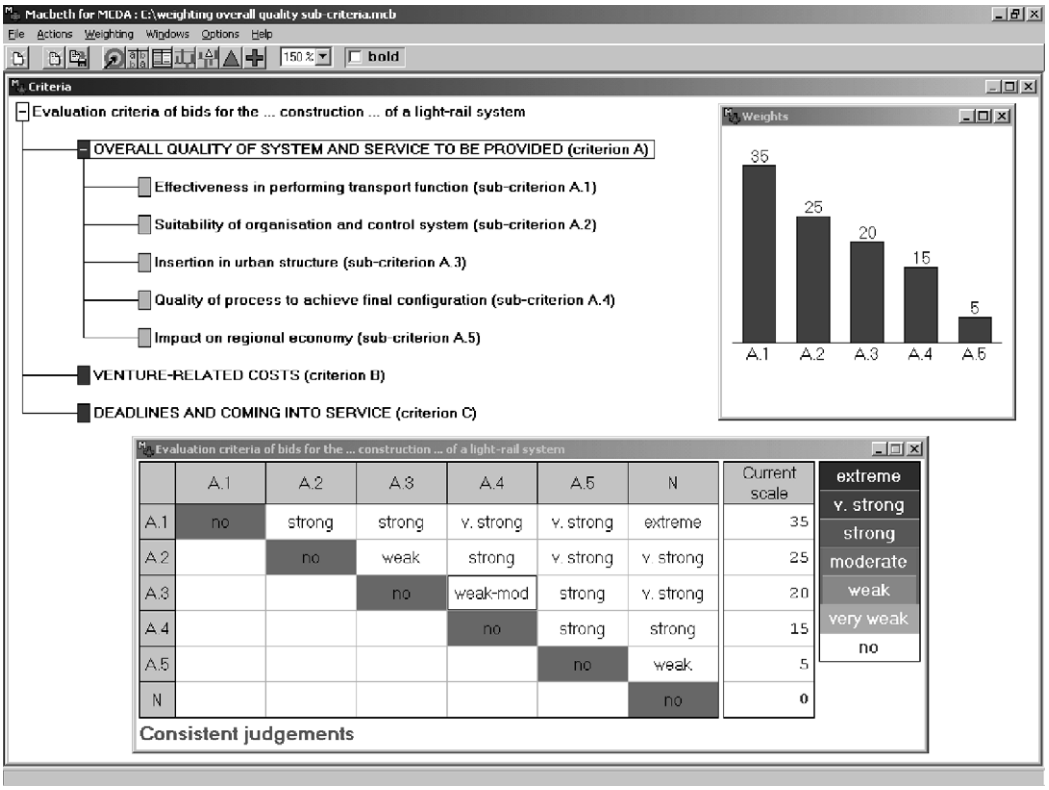
To determine their respective values, the application of the trade-off procedure would require partial value functions to have been previously defined. (Even if this had been the case, problems often arise in the assessment of trade-offs between qualitative impacts.) The swing weighting procedure could be used, by asking the Evaluation Committee: "Let the swing from neutral to good on the best ranked criterion values 100 (overall value units); what is the relative value of the swing from neutral to good on the criterion ranked second?" (and so on). However, from our practical experience, this numerical question is difficult to answer and the process of reconciling different numerical values for the same swing given individually by members of the group is a difficult task, even with the help of "visual analogue scales" [7, pp. 12–19]. Note that to take an average of these values lacks substantive meaningfulness and can infringe upon the ranking of the weights previously agreed.

Alternatively, the weighting procedure we use resorts to the question–answer protocol of the MACBETH approach.

3.2. MACBETH *weighting process: case study*

The process of the "International Public Call for Tenders to award the design, construction, equipment, financing and short-term operation, of a light-rail system in the Metropolitan Area of Porto", issued by Metro do Porto, S.A. in 1996, took place in three stages: first stage—pre-qualification, second stage—selection and third stage—negotiation. We will now examine the second stage, in which the evaluation criteria were the ones displayed in Fig. 2a. The respective weights were defined following the MACBETH approach.

To estimate weights with MACBETH, the decision-making group is only required to make qualitative judgements, for



	A.1	A.2	A.3	A.4	A.5	N	Current scale
A.1	no	strong	strong	v. strong	v. strong	extreme	35
A.2		no	weak	strong	v. strong	v. strong	25
A.3			no	weak-mod	strong	v. strong	20
A.4				no	strong	strong	15
A.5					no	weak	5
N						no	0

Fig. 3. Example of MACBETH weighting process—Metro do Porto S.A.

each pair of criteria j and j' with $k_j > k_{j'}$, of the difference in overall attractiveness between swinging from neutral to good in j instead of swinging from neutral to good in j'—that is, the difference of attractiveness between the fictitious bid b_j (good in j and neutral in the other criteria) and $b_{j'}$ (good in j' and neutral in the other criteria). The judgements are expressed by choosing one of the MACBETH semantic categories ("very weak", "weak", "moderate", "strong", "very strong", or "extreme" difference of attractiveness). Each judgement should reflect a collective view of difference in attractiveness. The judgements can be represented in a matrix (if two criteria have the same weight, "no" is introduced in the matrix), as the one displayed in Fig. 3 for the five sub-criteria (A.1–A.5) of the *overall quality of system and service to be provided* criterion (A). If there is judgemental hesitation or disagreement among group members about which category better reflects the difference of attractiveness, a discussion within the group should be launched. However, to avoid forcing unrealistic consensus, more than one category can be chosen. This was the case for the judgement "weak to moderate" in the cell (A3, A4) in the matrix of Fig. 3.

Notice that the last column of the matrix displayed in Fig. 3 is also filled in with judgements. This is due to the fact that at the time of this intervention (1996), we used to facilitate (or confirm) the initial ranking of the swings by asking the Evaluation Committee to judge the overall added attractiveness of the fictitious bid N ("neutral in all sub-criteria") brought about by improving its neutral impact to good in sub-criterion j ($j = 1, \ldots, n$)—that is, the overall difference of attractiveness between the fictitious bids b_j (good in j and neutral in other sub-criteria) and N. Our experience with the use of MACBETH to support group processes of weighting criteria in the last 7 years, has led us to conclude, empirically, that in general these initial questions are dispensable (except perhaps when the number of criteria is smaller than 4). Either way, it is up to the facilitators choose whether or not to fill in the last column (in this case only the order of the swings is considered—see example in [16]), according to the characteristics of each particular case with which they are faced. (MACBETH accepts all forms of incomplete matrices.)

The weighting process develops in one decision conference with the support of the MACBETH software. Each time a judgement is formulated by the decision-making group, the facilitator introduces it in the matrix, and the software automatically tests the consistency of all the judgements already formulated, points out eventual instances of inconsistency and, in such cases, gives suggestions to facilitate the revision of the judgements by the group in order to achieve consistency. The bar chart in Fig. 3 shows the numerical values (in terms of percentage) that MACBETH suggested to be assigned to the relative weights of the sub-criteria.

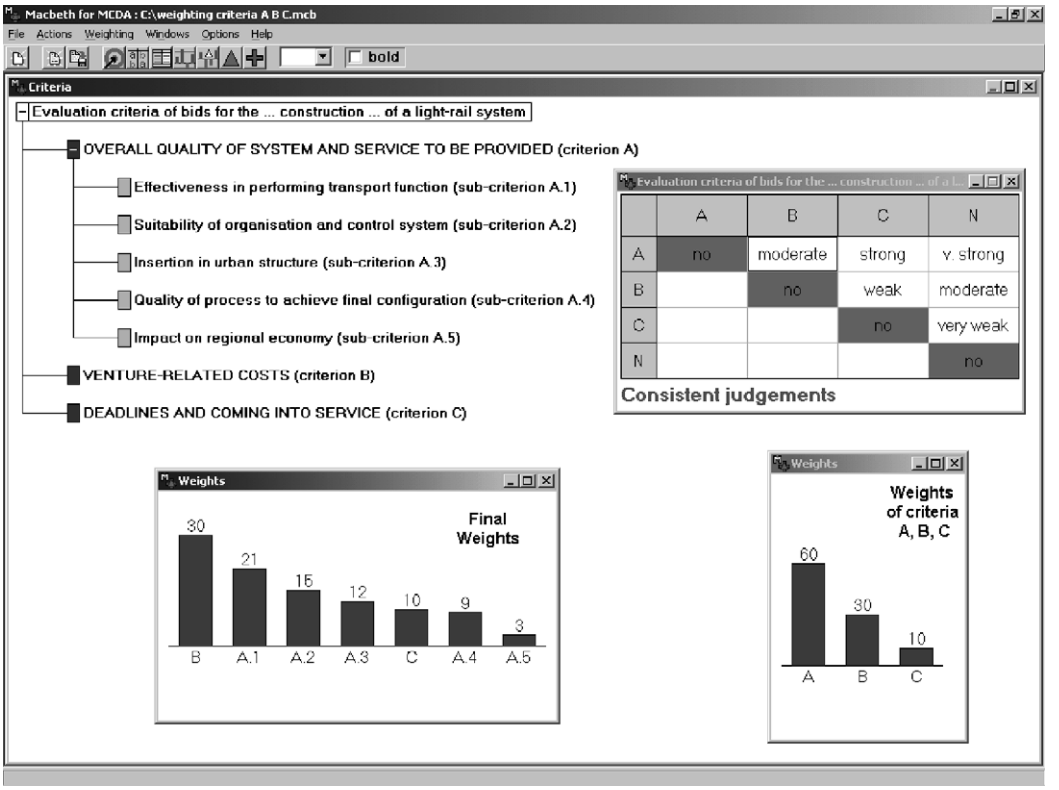


Fig. 4. Final weights (in percentages) of evaluation criteria and sub-criteria ("construction of a light-rail system" call for tenders—Metro do Porto S.A.)

As shown in Fig. 4, a similar process was followed to define relative weights to the criteria (A, B and C). The bottom left chart shows the final weights (in terms of percentage) for the criteria and sub-criteria as announced in the Evaluation Regulation (the values of the final weights for the sub-criteria A.1–A.5 were obtained by multiplying their "within criterion" weights, displayed in Fig. 3, by the weight of their parent criterion A).

As made obvious by the contents of this section, weighting criteria is a very sensitive and not-at-all trivial process, which does not always receive enough attention in the literature—an example is the short paragraph dedicated to the subject in [1, p. 368].

The definition of screening and evaluation criteria as well as their respective descriptors, levels of reference and relative weights are the minimal multicriteria information needed to draw up the Call for Tenders and the Evaluation Regulations.

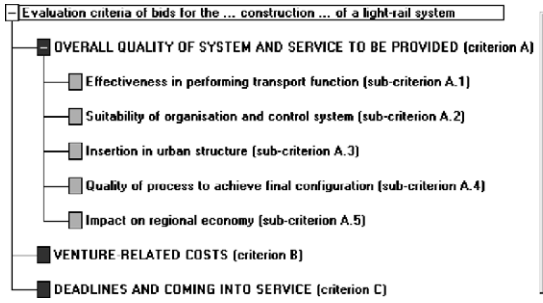
4. Evaluation phase: case study

4.1. Evaluation of bids according to criteria and sub-criteria

From the seven international groups that presented applications to Metro do Porto S.A., only four of them (G1, G2, G3 and G7) submitted bids to the second stage, designated 1.1, 1.2, 1.2' (variant proposal) and 1.3 presented by Group G1, 2.1 and 2.2 by Group G2, 3.1 and 3.2 by Group G3, and 7.1 and 7.1' (variant proposal) by Group G7, all respecting the screening criteria and therefore having been accepted.

The task of the Evaluation Committee (composed of five engineers, one economist and one jurist) was to evaluate these bids and recommend which two bidders should be selected to move on to the third stage—negotiation. As stated in the Evaluation Report, "{...} the Evaluation Committee started by adopting, as a key thread in its work methodology, an Evaluation Regulation for bid evaluation, that established the organic support and procedural as well as methodological framework to perform its functions. Obviously, this Regulation fully adopts the evaluation criteria established in the {...} Call for Tenders. In accordance with the {...} Evaluation Regulation, analysis of the bids was carried out by applying a Multicriteria Decision Aid Methodology".

In a series of decision conferences facilitated by the first author of this paper, the seven members of the Evaluation Committee (and only them) evaluated the (partial) attractiveness of the bids in each of the sub-criteria A.1–A.5 and criteria B and C. Each decision conference developed in two parts:



Actions rankings by criteria

Effectiveness	Suitability	Insertion	Process	Economy	Costs	Deadlines
good	good	3.1	good	3.1	good	good
3.1	1.1	3.2	3.1	7.1	1.2	1.2
1.1	1.3	1.2	3.2	7.1'	3.1	1.2'
1.3	1.2'	good	1.1	good	3.2	3.1
1.2'	7.1'	1.2'	1.3	1.2	1.2'	3.2
3.2	2.1	1.1	7.1	1.2'	2.2	1.1
7.1'	2.2	1.3	7.1'	1.1	1.3	1.3
neutral	1.2	2.1	neutral	1.3	2.1	neutral
1.2	3.1	2.2	1.2	2.2	neutral	2.1
7.1	3.2	neutral	1.2'	neutral	1.1	2.2
2.1	neutral	7.1	2.1	2.1	7.1	7.1
2.2	7.1	7.1'	2.2	3.2	7.1'	7.1'

Fig. 5. Results of the first part of the decision conferences: ranking the bids in each evaluation criterion or sub-criterion (in each column, bids in the same box were considered indifferent)—Metro do Porto S.A.

Impact on regional economy (sub-criterion A.5)

	3.1	7.1	7.1'	good	1.2	1.2'	1.1	1.3	2.2	neutral	2.1	3.2	Current scale
3.1	no	very weak	very weak	weak	weak	weak	weak	weak	v. strong	v. strong	v. strong	extreme	140.00
7.1		no	no	weak	weak	weak	weak	weak	v. strong	v. strong	v. strong	extreme	130.00
7.1'		no	no	weak	weak	weak	weak	weak	v. strong	v. strong	v. strong	extreme	130.00
good				no	no	no	very weak	very weak	strong	strong	strong	v. strong	**100.00**
1.2				no	no	no	very weak	very weak	strong	strong	strong	v. strong	100.00
1.2'				no	no	no	very weak	very weak	strong	strong	strong	v. strong	100.00
1.1							no	no	strong	strong	strong	v. strong	90.00
1.3							no	no	strong	strong	strong	v. strong	90.00
2.2									no	very weak	weak	strong	10.00
neutral										no	very weak	moderate	**0.00**
2.1											no	moderate	-10.00
3.2												no	-70.00

Consistent judgements

Fig. 6. MACBETH matrix of judgements for sub-criterion A.5—Metro do Porto S.A.

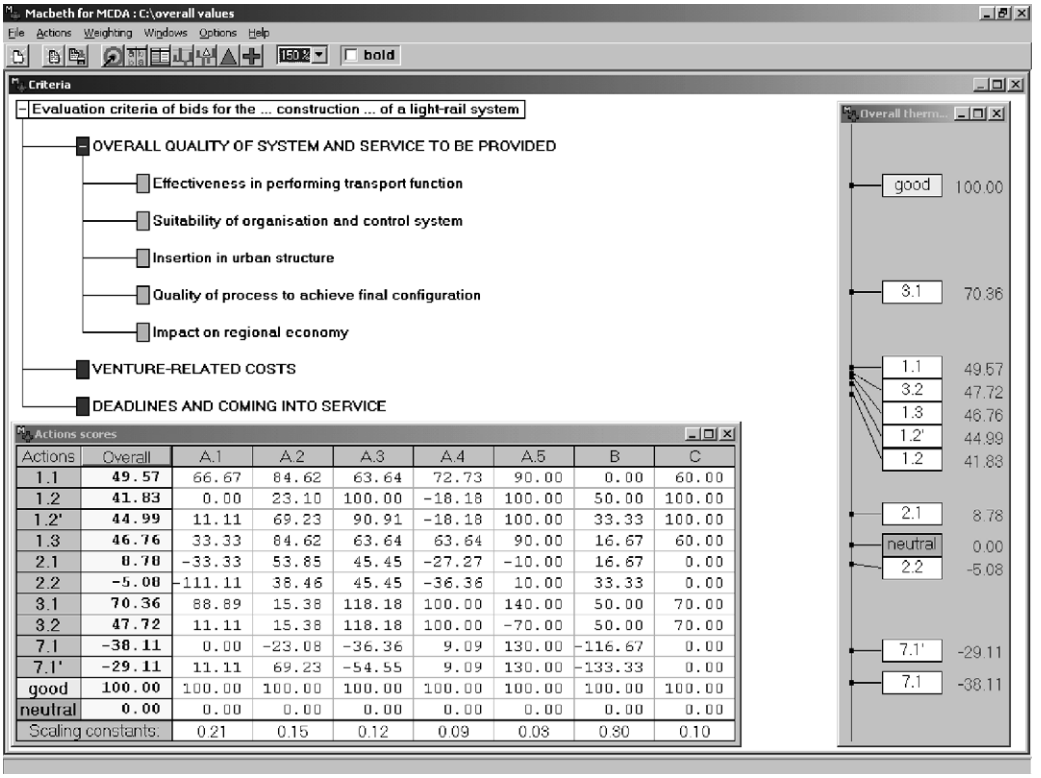
1. For each evaluation criterion or sub-criterion, the Evaluation Committee began by determining the impacts of the bids referring to the documents that constitute the bid submissions and basing themselves on detailed analysis reports of the bids drawn up by several external consultants. The bids were then classified as "very positive" (at least good), "positive" (at least neutral but less attractive than good) or "negative" (less attractive than neutral) and ordered by descending order of attractiveness (Fig. 5 summarises the rankings established in the first parts of the various decision conferences).
2. In the second part, the Evaluation Committee performed group judgements regarding the difference of (partial) attractiveness between each two bids, including the references "good" and "neutral". For example, Fig. 6 shows the consistent matrix of qualitative judgements on sub-criterion A.5 and the respective value scale in resulting from the application of MACBETH.

Once the partial evaluation phase was completed, the additive model was applied to aggregate the partial values of the bids (summarised in the bottom left table in Fig. 7) to calculate the overall values which reflected the overall attractiveness of the bids. As can be seen in the overall value "thermometer" displayed in Fig. 7, no bid's result was very positive (better than good) and three bids turned out to be negative (worst than neutral). The two most attractive bids, 3.1 and 1.1, are from different bidders, Groups G3 and G1, respectively. Moreover, the overall values of all G3's and G1's bids are largely better than the overall values of all G2's and G7's bids (2.1, 2.2, 7.1, 7.1'). For that reason, the Committee formed an initial idea of its selection recommendation, which is then submitted to sensitivity and robustness analyses.

4.2. Sensitivity and robustness analyses

4.2.1. Modifying partial values

The Evaluation Committee decided to proceed to the validation of the robustness of the final ranking of the bids, in terms of overall attractiveness, according to the following



Actions scores

Actions	Overall	A.1	A.2	A.3	A.4	A.5	B	C
1.1	49.57	66.67	84.62	63.64	72.73	90.00	0.00	60.00
1.2	41.83	0.00	23.10	100.00	-18.18	100.00	50.00	100.00
1.2'	44.99	11.11	69.23	90.91	-18.18	100.00	33.33	100.00
1.3	46.76	33.33	84.62	63.64	63.64	90.00	16.67	60.00
2.1	8.78	-33.33	53.85	45.45	-27.27	-10.00	16.67	0.00
2.2	-5.08	-111.11	38.46	45.45	-36.36	10.00	33.33	0.00
3.1	70.36	88.89	15.38	118.18	100.00	140.00	50.00	70.00
3.2	47.72	11.11	15.38	118.18	100.00	-70.00	50.00	70.00
7.1	-38.11	0.00	-23.08	-36.36	9.09	130.00	-116.67	0.00
7.1'	-29.11	11.11	69.23	-54.55	9.09	130.00	-133.33	0.00
good	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
neutral	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Scaling constants:		0.21	0.15	0.12	0.09	0.08	0.80	0.10

Fig. 7. Evaluation outputs—Metro do Porto S.A.

four scenarios of partial values modifications:

A. Scenario in which partial evaluation of bids in criterion A (*Quality*) is conducted considering equal relative weights for the five sub-criteria A.1–A.5 (limit situation respecting, lato sensu, the order of "importance" of the sub-criteria).
B. Scenario in which partial evaluation of bids in criterion B (*Costs*) is conducted considering a linear value function in the indicator *NPV of proposed global price* (ignoring all other cost indicators).
C. Scenario in which partial evaluation of bids in criterion C (*Deadlines*) is conducted considering a linear value function in the indicator *average deadline to come into service* (ignoring the delay between awarding and consignment).
D. Scenario in which partial evaluation of bids in criteria A, B and C is conducted considering scenarios A, B and C simultaneously.

Analysing the results of the application of the additive model in each of the four scenarios, the Committee concluded, as stated in the Evaluation Report, that: "in the first three scenarios above, bids 3.1 and 1.1 (in this order) are always the best, except for the scenario D (conjunction of the previous ones). However, even in this scenario the fundamental general conclusion is maintained: there is always at least one bid from Group 3 and one from Group 1, that are better classified in terms of overall value than any bid from Groups 2 and 7".

4.2.2. Modifying relative weights

The most popular type of sensitivity analysis of weights consists of analysing the changes that may occur in the global ordering of the bids when the relative weight of a given criterion (or sub-criterion) is modified, keeping the proportion among other weights. For example, Fig. 8 shows a sensitivity analysis graph for the weight of the sub-criterion A.2.

Besides the classic sensitivity analysis, it is very interesting to analyse the robustness of the results by making several weights vary at the same time, but respecting the order of the weights defined in the call for tenders. The first version of the software PROBE (*P*reference *Rob*ustness *E*valuation—cf. [17]; see detailed example in [16]) was designed precisely for this purpose at the time of our intervention in Metro Porto S.A. PROBE uses the concept of *additive dominance* [18]: a bid x is said to dominate additively a bid y, for a certain weights ranking, if the difference between the overall values of x and y is always positive for any vector of weights that respects the ranking of the weights. Fig. 9 shows the "table of dominances" for the case in which only the ranking of the weights of criteria A, B and C (by this

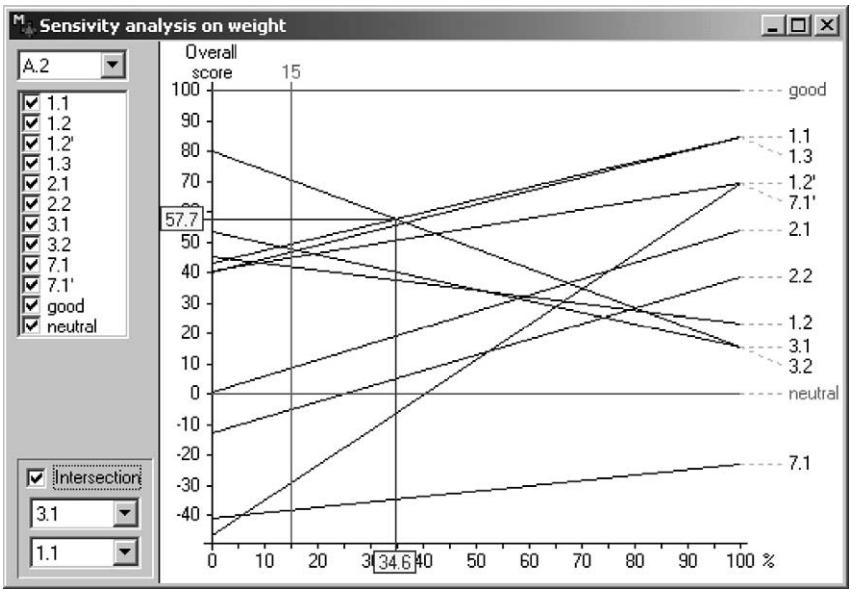


Fig. 8. Sensitivity analysis on the weight of sub-criterion A.2 (the vertical line represents the current value of the weight; 34.6 is the value of the weight for which the two best bids would be indifferently the best ones).

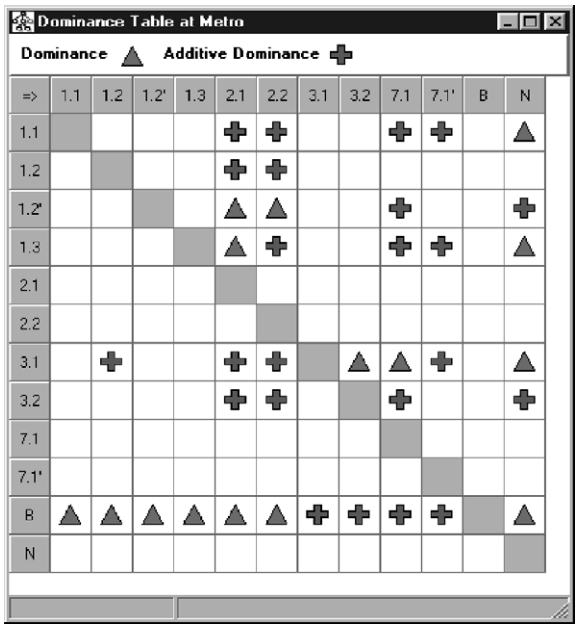


Fig. 9. Robustness analysis respecting simultaneously the ranking of the weights of the three criteria and the ranking of the weights of the five sub-criteria.

order) and the ranking of the weights of sub-criteria A.1–A.5 (by this order) are taken into consideration simultaneously (note this is an input significantly different from a complete ranking of the weights). As can be seen, the bids from group G2 (2.1 and 2.2) are dominated by all the bids from groups G1 and G3. Additionally, the bids from group G7 (7.1 and 7.1') are dominated by at least one bid from group G1 and one from G3. Additionally, no bid is dominated by the bids of G2 and G7. Since the objective of the second stage of the call for tenders was to select two groups for later negotiation, the above conclusions robustly confirm the selection of groups G1 and G3. The construction of a requisite model was achieved.

After the development of the extensive sensitivity and robustness analyses described above, the Evaluation Committee used all of this information as basis for its final decision. As stated in the Evaluation Report of the second stage, the final recommendation of the Evaluation Committee was to select groups G1 and G3 for the third stage of the process (negotiation).

5. Discussion and conclusion

5.1. A posteriori analysis of the case study

The following comment of João L. Porto, CEO of Metro do Porto S.A. at the time of our intervention, expresses the decision-makers' point of view regarding the use of this kind of methodology in a complex evaluation process: "When we were first introduced to the method (MACBETH), we recognise it was not easy to comprehend. However, after this initial effort to understand the logic behind its application, it became simple. We (the evaluators) had only to pairwise compare the bids in each criterion separately, and the logic was the same to estimate weights. In spite of

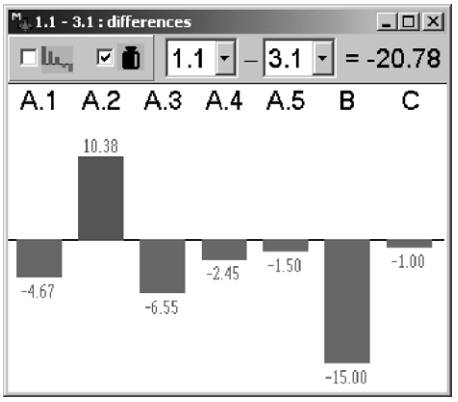


Fig. 10. Weighted value differences between bids 1.1 and 3.1 at the end of the second stage—Metro Porto S.A.

the demands in terms of reasoning and time, its use added value to the decision process". (Adapted from the original in French [19].)

It is worthwhile to report that, in sequence with the decision taken at second stage of the Metro Porto S.A process, the two bidders selected (G1 and G3) were invited to improve their best bids, 1.1 and 3.1, respectively, in any of the evaluation criteria, provided that their partial values did not decrease in any of them, a condition that was respected by the two groups. After the negotiation, the two bids were re-analysed and re-evaluated taking into consideration the ameliorations introduced. At the end, the contract was awarded to Group G1, which means that G1 succeeded in improving its bid enough to overcome its initial disadvantage towards G3 (see Fig. 10).

Since 1996, the robustness analysis reasoning and the operationalisation of the concept of additive dominance have been deeply extended in the MACBETH approach. The M-MACBETH software (available at http://www.umh.ac.be/∼smq) now allows users to perform dominance analyses for many situations of poor preference information other than weight rankings, and not only on the weights but also on the partial values (cf. [5]). M-MACBETH displays a "control panel" for robustness analysis shown in Fig. 11, that demonstrates the use of this feature for cases in which the constraints for additive dominance are the ones implied by the matrices of judgements in the evaluation criteria and the ranking of their weights. The output table confirms the decision taken by the Evaluation Committee in 1996. The cross-hatched cells correspond to "incomparability" situations: if the cells (x, y) and (y, x) are cross-hatched, then there exist some numerical values, compatible with the preference information taken into account, for which the output of the additive model is "x globally better than y" and other ones (also compatible with the same information) for which the output is "y globally better than x" (which is a meaningless result).

5.2. Conclusion

Each process of bid evaluation in which we have participated has proved to be a distinct case, with its own problems and specific characteristics, which suggests prudence in the generalisation and extrapolation of conclusions to other contexts. However, it is safe to point out as a common feature that each Evaluation Committee displayed, in regards to our multicriteria decision aid methodology:

- agreement with the assumptions of its application;
- assimilation of its essential theoretical principles;
- natural agreement with the sequence of phases proposed to construct the evaluation model;
- satisfaction with the contribution provided to ensure coherence and formal correctness in procedures;
- approval of the MACBETH qualitative judgement process, and the freedom with which it provided them to verbalise their own values. In the words of the Evaluation Committee of the Porto case in its final report: "the method has never been, throughout the whole process, a factor constraining decision-making by the Evaluation Committee, which has always been exclusively responsible for formulating the necessary value judgements".

Finally, there are two key conclusions derived from reflecting upon our interventions as facilitators and analysts in public sector processes such as bid evaluation in public call for tenders as well as in other multicriteria decision analyses developed for private companies. On the one hand, in both types of frameworks we have been adopting a constructive decision aid attitude, closed to what [20] calls "process consultation": the core of our intervention has been to contribute to the process of developing a common language for stimulating group interaction and debate, using the technology as an instrument to assist us in helping our clients to increase their knowledge about their problems and to find their own solutions for them. On the other hand, interventions on bid evaluation in public call for tenders are much more demanding, both from the facilitators' and the clients' perspectives: it is not enough to help the decision-making group to achieve an agreement on a decision, it is also necessary to help them justify it in a clear and unambiguous way. As a matter of fact, it is not easy to deal with the tension provoked a priori in Evaluation Committees by the mere hypothesis of potential future (legal) disputes about the integrity of their judgements and decisions. A public call for tenders put out by Câmara Municipal do Funchal (Municipality of Funchal, in Madeira Islands) has been recently contested in court by the bidder ranked the second best by the Evaluation Committee. It is interesting to report that, in his sentence, the Judge explicitly emphasised the robustness and correctness of the multicriteria methodology applied in the bid evaluation (under the facilitation of the first author and

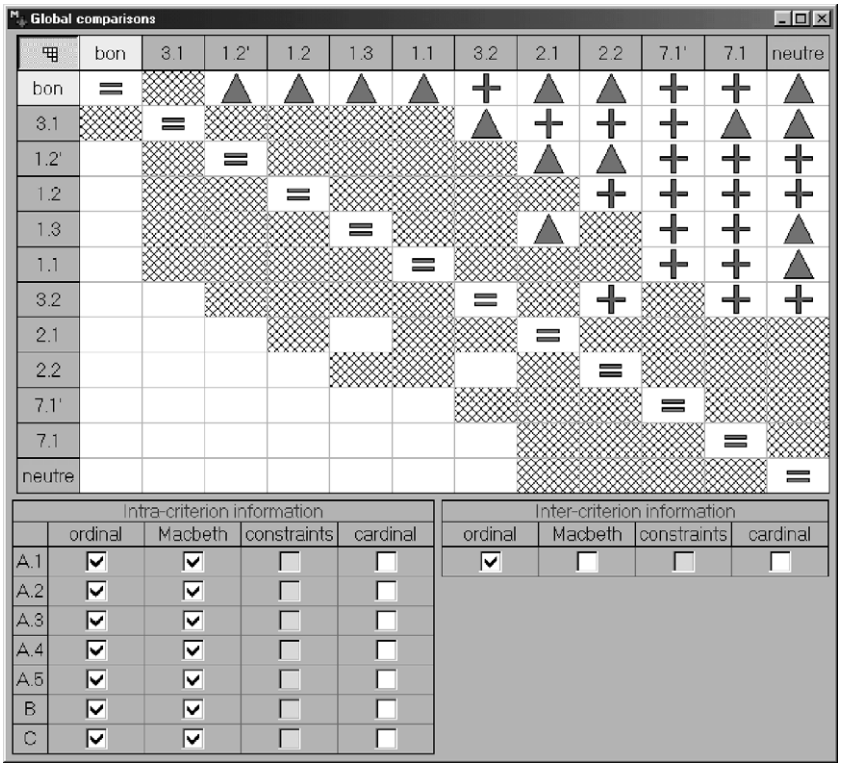


Fig. 11. A posteriori robustness analysis performed with the M-MACBETH for MCDA software.

another colleague), but agreed with the argument presented by the contester's lawyers that the participation of the two consultants in the meetings of the Evaluation Committee would require, from a legal point of view, that they should have been formally appointed by the Mayor as members of the Evaluation Committee, which had not been the case!

Acknowledgements

The authors would like to thank Metro do Porto, S.A. and Metropolitano de Lisboa, E.P. for the authorisation to use selected information from their calls for tenders, and all the members of Evaluation Committees from *Porto*, *Lisbon and Mondego Metro Companies*, *Lisbon and Funchal Municipalities*, *Lisbon Gas Company*, *Portuguese Water Resources Institute and Companies*, *Porto and Portuguese Industrial Associations*, *Portuguese Roads Institute*, *inter alia* for having offered the first two authors the opportunity to carry out an intellectually gratifying process-consultation activity that has resulted in the methodological development presented in this article. Finally, the collaboration of José A. Antunes Ferreira in many consulting applications and refinements of the methodology must also be emphasised.

References

[1] Liu SL, Wang SY, Lai KK. Multiple criteria decision making models for competitive bidding. In: Shi Y, Zeleny M, editors. New frontiers of decision making for the information technology era. London: World Scientific, 2000. p. 349–72.

[2] Phillips LD. Decision analysis for group decision support. In: Eden C, Radford J, editors. Tracking strategic problems. London: Sage, 1990. p. 142–50.

[3] Bana e Costa CA, Vansnick JC. Applications of the MACBETH approach in the framework of an additive aggregation model. Journal of Multi-Criteria Decision Analysis 1997;6(2):107–14.

[4] Bana e Costa CA, Vansnick JC. The MACBETH approach: basic ideas, software and an application. In: Meskens N, Roubens M, editors. Advances in decision analysis. Dordrecht: Kluwer Academic Publishers, 1999. p. 131–57.

[5] De Corte JM, Vansnick JC, Bana e Costa CA. Progressive exploration of preferential information in multicriteria decision aid. Working Paper 17/01, Centre of Management Studies (CEG-IST), Instituto Superior Técnico, Lisbon, 2001.

[6] Phillips LD. A theory of requisite decision models. Acta Psychologica 1984;56:29–48.

[7] Belton V. Multi-criteria problem structuring and analysis in a value theory framework. In: Gal T, Stewart T, Hanne T, editors. Multicriteria decision making, advances in MCDM—models, algorithms, theory, and applications. Dordrecht: Kluwer Academic Publishers, 1999. p. 12–132.

[8] Eden C. Cognitive mapping: a review. European Journal of Operational Research 1988;36:1–13.
[9] Bana e Costa CA, Ensslin L, Corrêa EC, Vansnick JC. Mapping critical factors for firm sustainable survival: a case-study in the Brazilian textile industry. In: Kersten G, Mikolajuk Z, Rais M, Yeh A, editors. Decision support systems for sustainable development in developing countries. Dordrecht: Kluwer Academic Publishers, 1999.
[10] Roy B. The outranking approach and the foundations of ELECTRE methods. In: Bana e Costa CA, editor. Readings in multiple criteria decision aid. Berlin: Springer, 1990. p. 155–83.
[11] Keeney RL. Value-focused thinking: a path to creative decision making. Cambridge, MA: Harvard University Press, 1992.
[12] Keeney RL, Raiffa H. Decisions with multiple objectives: preferences and value tradeoffs. New York: Wiley, 1976.
[13] Von Winterfeldt D, Edwards W. Decision analysis and behavioral research. New York: Cambridge University Press, 1986.
[14] Goodwin P, Wright G. Decision analysis for management judgement, 2nd ed. New York: Wiley, 1998.
[15] Von Nitzsch R, Weber M. The effect of attribute ranges on weights in multiattribute utility measurements. Management Science 1993;38(8):937–43.
[16] Bana e Costa CA. The use of multicriteria decision analysis to support the search for less conflicting policy options in a multi-actor context: case-study. Journal of Multi-Criteria Decision Analysis 2001;10(2):111–25.
[17] Lourenço J, Bana e Costa CA. PROBE for Windows user manual. Lisbon: Cised Consultores, Lda., 1998.
[18] Bana e Costa CA, Vincke P. Measuring credibility of compensatory preference statements when trade-offs are interval determined. Theory and Decision 1995;39(2): 127–55.
[19] Porto JL. Analyse multicritère dans le cadre des appels d'offres pour la construction de travaux publics et privés : le cas du Métro de Porto au Portugal. Newsletter of the European working group multicriteria aid for decisions, Series 2, No. 15, 1999. p. 1–2.
[20] Schein EH. Process consultation revisited: building up the helping relationship. Reading, MA: Addison-Wesley, 1999.