Carlos A. Bana e Costa, João C Lourenço, Manuel P. Chagas and João C. Bana e Costa,

Development of reusable bid evaluation models for the Portugese Electric Transmission Company

Working paper

Development of Reusable Bid Evaluation Models for the Portuguese Electric Transmission Company

Carlos A. Bana e Costa
CEG-IST, Centre for Management Studies of IST, Technical University of Lisbon, Av. Rovisco Pais, 1049-001 Lisbon, Portugal, carlosbana@ist.utl.pt
Department of Management - Operational Research Group, London School of Economics, Houghton Street, London WC2A 2AE, UK, c.bana@lse.ac.uk
João C. Lourenço
CEG-IST, Centre for Management Studies of IST, Technical University of Lisbon, Av. Rovisco Pais, 1049-001 Lisbon, Portugal, joao.lourenco@ist.utl.pt
Manuel P. Chagas
University of Chicago, Graduate School of Business, Full-Time MBA Program, 5807 South Woodlawn Avenue, Chicago, IL 60637, MChagas@chicagogsb.edu
João C. Bana e Costa
BANA Consulting, Lda., R. Prof. Bento Jesus Caraça, 33, 1600-600 Lisbon, Portugal, joao@bana-consulting.pt

Bid evaluation is the process of selecting a contractor from a number of bidders. The decision analysis models currently in use at the Portuguese Electric Transmission Company (REN) to evaluate bids were developed through a decision conferencing process supported by the MACBETH multicriteria approach and software. This paper presents the various components of this interactive socio-technical process. Given the number of contracts awarded by REN each year, it was crucial that the models be reusable in similar calls for tenders; this required substantial care in structuring the criteria, with a focus on constructed scales, and building value function models based on qualitative pairwise comparison judgments of difference in attractiveness. Also of particular interest is the approach for weighing benefits against costs.

Key words: bid evaluation; constructed scales; decision conferencing; MACBETH; model structuring; multicriteria weighting

1. Introduction

The Portuguese Electric Transmission Company (REN - Rede Eléctrica Nacional, S.A.) is the public organization responsible for transmission of electricity in Portugal. To accomplish its mission, REN regularly awards contracts for various external services, supplies and projects, that are in compliance with national and European Union regulations relating to public calls for tenders (namely Directive 2004/18/EC of the European Parliament and of the Council, 31 March 2004, on the coordination of procedures for the award of public works contracts, public supply contracts and public service contracts). The contracting process begins with an invitation to tender that is sent to a small number of pre-qualified service providers or suppliers. "The process of selecting a contractor from a number of bidders is called bid evaluation" (Mustafa and Ryan 1990). Evaluating the attractiveness of the bids against the multiple criteria announced in the invitation to tender is a key step in this process. In April 2005, Henrique Gomes, from the REN Board of Directors, asked the first author to analyze

the bid evaluation system used by REN and, depending on the outcome of the analysis, either correct it or develop a new system. The purpose of this paper is to present our subsequent socio-technical intervention at REN, addressing social aspects of decision conferencing combined with technical components of building a multicriteria bid evaluation model, using MACBETH (Measuring Attractiveness by a Categorical Based Evaluation Technique). MACBETH is an interactive multicriteria decision analysis approach used to build a quantitative (numerical) value model based on qualitative (non-numerical) pairwise comparison judgments (Bana e Costa and Vansnick, 1994 and 1999, Bana e Costa et al. 2005a). A brief technical overview of MACBETH is presented in Section 2.3.

REN publishes invitations to tender when constructing new power lines, upgrading existing lines, constructing electrical facilities, purchasing equipment, acquiring power substations systems, or developing engineering projects. Each of these types of initiatives requires that technical staff from different departments be included in the evaluation of bids. Therefore, the authors recommended that a decision conferencing process (Phillips and Phillips 1993, Phillips 2007, Phillips and Bana e Costa 2007), which requires the involvement of participants from all relevant departments, be used to analyze the current evaluation system in detail as a basis for the development of a new model(s), as it would ensure broad representation. A Working Group (WG) was formed: it consisted of six departmental executive managers headed by Henrique Gomes. The group discussions were guided by a process consultant facilitator (Schein 1999), who was assisted by two decision analysts mainly responsible for operating M-MACBETH (Bana e Costa et al. 2005b), a software application that implements the MACBETH approach. M-MACBETH supported the development and use of the new bid evaluation system.

In our review of the literature, we have found several technical presentations of bid evaluation methods, but almost no detailed descriptions of real-world socio-technical bid evaluation group processes. With these findings in mind, we deliberately structured this paper to closely follow the actual development of the REN decision conferencing process (Sections 3 through 6). Also, the figures include content generated on the spot, to emphasize the importance of visual interactivity in constructive value modeling. We share the perspective that decision analysts "should function not as archaeologists, carefully uncovering what is there, but as architects, working to build a defensible expression of value" (Gregory et al. 1993, p. 179).

The process started with a kick-off meeting with the WG (Section 3) during which the model building tasks were identified, and agreed upon, as depicted in Figure 1. Before describing the process in detail, we discuss in Section 2.1 key contextual and methodological

issues that drove the design of our three-day decision aid intervention at REN, during which two models were developed. To preserve industrial confidentiality, some of the data presented have been altered or disguised.

2. Contextual and Methodological Issues

2.1 Selecting an Approach to Build Reusable Bid Evaluation Models

The Most Economically Advantageous Bid

Directive 2004/18/EC dictates that when multiple criteria, rather than just price, are used to evaluate bids, as is the case in REN, the award must be made to "the tender most economically advantageous to the contracting authority" (art. 56). From a technical perspective, this requires value trade-offs across criteria and invites the development of a multicriteria additive aggregation model: value scores (directly or indirectly) assigned to each bid are multiplied by the respective weights assigned to the criteria and those products are summed across all of the criteria. The bid with the highest overall value score is specified as the 'most economically advantageous' one.

Bid Evaluation and Decision Analysis Procedures

Among the various bid evaluation procedures described in the literature (see surveys in Holt 1998 and Liu et al. 2000) the use of multicriteria value methods proposed in decision analysis (von Winterfeldt and Edwards 1986, Kirkwood 1997, Belton and Stewart 2002) undoubtedly respect those legal requirements while also being theoretically sound, which is not necessarily the case with many other ad-hoc bid evaluation methods. This paper assumes that the reader is familiar with the theoretical foundations and traditional procedures for the construction of value models (cf. Krantz et al. 1971, Keeney and Raiffa 1976, French 1986, Dyer and Sarin 1979). In practice, assigning value scores to bids and weights to criteria requires the assessment of value judgments from the evaluators. These judgments can be quantitative or qualitative, depending on the specific scoring and weighting procedures used. Among the most widely used techniques, direct rating and swing weighting (Edwards and Barron 1994) require numerical estimations from the evaluators, while the bisection (or middle point splitting) and trade-off techniques (Keeney and Raiffa 1976) are based on indifference judgments involving at least three elements in each judgment. Several applications involving numerical or indifference assessments can be found in different bid evaluation contexts. (See, for example, Sarin et al. 1978, Dyer and Lorber 1982, Belton 1985, Buede and Bresnick 1992, Ewing, Jr. et al. 2006.) In contrast, the MACBETH approach requires qualitative pairwise comparison judgments of difference in attractiveness (value), therefore involving only two elements in each judgment, to help an individual or

group evaluator to score options on each criterion and to weight criteria (Bana e Costa and Chagas 2004). The original research on MACBETH was carried out in the early 90's (Bana e Costa and Vansnick 1994) and has since been extensively applied in various contexts, one of which has been bid evaluation in public international calls for tenders (Bana e Costa and Vansnick 1997, Bana e Costa et al. 2002). In particular, several MACBETH value models were constructed and successfully used for human resource and supplier evaluation and management (Oliveira and Lourenço 2002) in the Lisbon Gas Company, when Henrique Gomes was a director, before moving to REN. He was, therefore, aligned with, and sympathetic to, the qualitative assessment of values and saw MACBETH as potentially useful at REN, too, because, in his opinion, there was a need to refocus the REN staff from their current biased numerical assessment framework. In fact, their ad-hoc scoring and weighting approach violated basic theoretic conditions for additive value modeling, as is often the case with, unfortunately popular, "point systems" (Hatush and Skitmore 1997). We, therefore, proposed a MACBETH socio-technical process, which was accepted by the REN Board of Directors and subsequently by the WG.

Direct and Indirect Bid Evaluation

At the level of each criterion, a bid evaluation model can be built following one of three paths: bids can be directly rated, or compared to one another (Dyer and Lorber 1982, Belton 1985, Mustafa and Ryan 1990, Bana e Costa et al. 2002); bids can be compared indirectly through a value or utility function (Ewing, Jr. et al. 2006, Pongpeng and Liston 2003) previously constructed upon a defined attribute; finally, bids can be compared through a hybrid of the aforementioned approaches, e.g., the bids might be compared directly on some quality criteria and indirectly on other criteria such as cost or deadline through previously assessed value functions. Each of these frameworks has advantages and drawbacks. The direct approach allows one to capture all of the characteristics of a set of bids since bid performances are known before the evaluation model is constructed, which is a clear advantage of this approach; the indirect approach may not detect every relevant characteristic that could have been used to evaluate a set of bids, because the bids are not known when the evaluation model is constructed: judiciously structuring the model (Section 4) is therefore essential to building a sound value (or utility) function model. The time and effort required to construct a bid evaluation model when using the direct approach is contingent on the number of bids evaluated, making the approach unattractive when facing a large number of bids, while the effort required for the indirect approach is unrelated to the number of bids under evaluation, which makes it more appealing when dealing with a large number of bids.

In addition, because the direct approach is dependent upon the bids themselves, the models built through this approach cannot be reused. The indirect approach, on the other hand, which creates a model that reflects value judgments regarding criteria irrespective of the specific bids, can be used time and again in similar evaluation processes. Therefore, the direct approach seems to be most adequate for one-time bid evaluation procedures in which few bids are evaluated, while the indirect approach is more appropriate when repeated evaluations, with the same value system, are required, or large numbers of bids are evaluated following a similar line of reasoning. Being aware of the weaknesses and strengths of both approaches helps one to choose the best framework for a given problem or to build a hybrid system using the best features of each approach.

Given the requirements of the REN Board (Section 3) that the evaluation system be built before bids are known and be reusable in similar calls for tenders, and that REN engineers should appraise but not score bids, the indirect path was followed to construct additive value function models that could, in the future, be reused to evaluate the bids of REN suppliers.

2.2 Model Building Tasks

Methodologically, the model building process used in this application can be described as a package of activities developed during decision conferences, grouped in Figure 1 into three main phases of analysis: structuring, evaluation, and testing the model requisiteness. The methodological basis for this model building process is now presented in more detail.

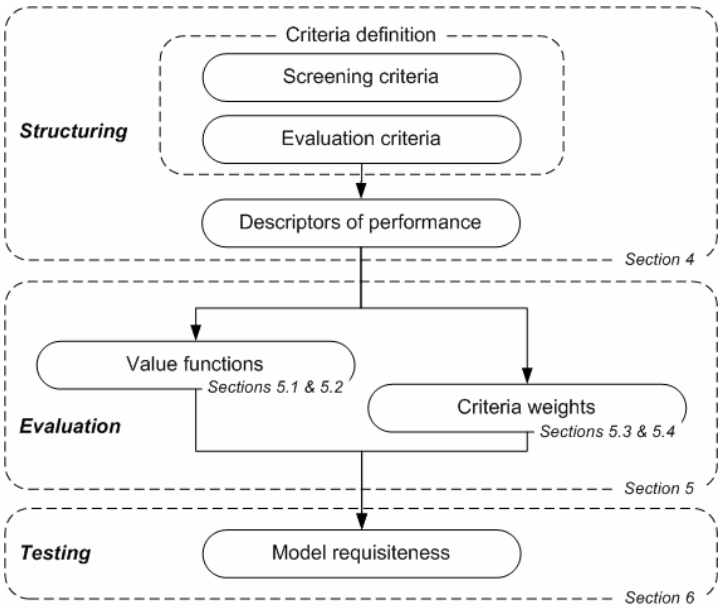


Figure 1. Model building tasks. The section where each task is presented in more detail is noted. Although the tasks shown are presented in a sequence it is possible to go back at any time to redefine or adjust what was previously done.

Structuring Activities: Defining and Operationalizing Criteria

The role of the facilitator during the interactive learning process of model-structuring consists of stimulating the reflection of the participants to, progressively, identify criteria, separating bid screening criteria (i.e., compulsory requisites to be respected by all proposals) from their evaluation criteria. Screening criteria, which are focused on bidders' capabilities (like technical experience or financial stability) rather than on their specific bids, are not legally acceptable for bid comparison; bidder screening criteria should only be considered for suppliers' prequalification or shortlisting. It can be argued that if both bidder and bid screening are based upon significantly demanding requisites, then the contract should be awarded to the lower price bidder. For instance, Hatush and Skitmore (1998) critically observe that: "By far the most frequently used method of selecting construction contractors is competitive bidding, in which the lowest bidder is awarded the contract." (p. 105). In their review of the contractor selection practice in the U.K., Holt et al. (1995) reinforce this view:

> Emphasis presently directed towards encouraging lower bid should be redirected towards establishing contractor ability for achieving client satisfaction (project performance, quality of completed product, etc.). The fundamental rationale behind competitive tendering is free market competition, i.e. genuine competition should achieve best value for money for the client; this has been the underlying philosophy of contractor selection for hundreds of years. … In 1979, the Institute of Building pointed out that shortcomings occurred in both the public and private sectors which resulted in clients having difficulty achieving best value for money. … Whatever the method, it is suggested that all selection pivots upon three criteria, namely: time (product completed as soon as possible); cost (product at the lowest possible price); and quality (product to be of the highest possible standard). (Holt et al. 1995, p. 553)

In many cases, these three criteria (time, cost and quality) appear immediately after the root node of a value tree of criteria and sub-criteria (see examples in Bana e Costa et al. 2002). However, 'minimize time' and 'maximize quality' can also be viewed as forming a broader objective 'maximize benefit', to be traded-off against 'minimize cost' at a last stage of analysis (see Section 5.4).

These or other evaluation criteria should be carefully selected; no additional criteria can legally be added after the bids are submitted. Moreover, since bids will be compared and scored in terms of their relative attractiveness with respect to each criterion individually, each criterion must be an independent evaluation axis. This is the reason why several means

objectives are often clustered into a single criterion rather than analyzed as independent sub-criteria (see Section 4). Desirable properties for a good set of "fundamental objectives" (Keeney 1992 and 2007) can be directly adopted for a family of bid evaluation criteria. The family of criteria should be consensual in the WG, and therefore exhaustive, but also concise and non-redundant. In addition, each criterion should be specific and understandable by the REN engineers to make it operational for the appraisal of bid performances.

A key task in building an indirect-evaluation model consists of associating with each criterion a (qualitative or quantitative) measure of the extent to which the criterion can be satisfied. This measure has been called an "attribute" in decision analysis, since Keeney and Raiffa (1976), although von Winterfeldt and Edwards (1986, p. 38) prefer "value dimension" and Kirkwood (1997, p. 24) uses "evaluation measure (also called measure of effectiveness, attribute or metric)." We prefer to call it a "descriptor of performance" in the context of bid evaluation (Bana e Costa et al. 2002) to avoid the common misinterpretation of the notion of an attribute as a criterion, characteristic or quality – as when one accepts as possible an "objective linked to one or more operational attributes" in "a traditional application of multiattribute value/utility analysis" (Butler et al. 2006, p. 100). Whatever the designation, the natural, proxy and constructed performance appraisal measures (see, for example, Kirkwood 1997, Keeney 1992, 2007) are all used in the bid evaluation context. The aim is always to describe bid performance on a criterion as objectively as possible, because the more objectively performance is appraised, the better understood (less ambiguous) and therefore the more accepted (less controversial) the evaluation model will be. In addition, non-desirable performance levels are defined as screening criteria, therefore restricting the range of performance on a criterion to a plausible range, from a most attractive to a least attractive level of performance. Specifically in the REN case, constructed scales were developed for all of the benefit criteria (see Section 4), following a systematic procedure proposed by Bana e Costa and Beinat (2005) to construct multidimensional scales for criteria that cluster several intertwined dimensions (see Table 1). For a large number of dimensions or levels, a full analysis of all possible combinations would be impracticable, and one can therefore opt for a few reference performance profiles covering the full range of the performance scale (Bana e Costa et al. (2002) present a technique to guide the definition of such reference levels; other suggestions can be found in Beinat (1997) and Belton and Stewart (2002, p. 130)).

Finally, whatever the descriptor of performance, our experience has revealed that expending the effort required to identify what is a good (unquestionably attractive) performance and a neutral (neither attractive nor unattractive) performance contributes

significantly to the intelligibility of the respective criterion. In particular, when developing constructed performance scales, starting by defining 'neutral' and 'good' levels can significantly facilitate the task (see Section 4).

Table 1. A procedure to develop a multidimensional constructed scale (source: Adapted from Bana e Costa and Beinat 2005, p. 23).

Basic steps	What to do	Comments
Step 1	Define a discrete set of performance levels in terms of each of the component dimensions.	This may require making the quantitative dimensions discrete.
Step 2	Establish all possible combinations of the levels of the various dimensions.	Developing all possible multidimensional combinations can be called a "factorial design" (Barron and Person, 1979); this can be facilitated using tables or trees.
Step 3	Eliminate infeasible (implausible) combinations.	
Step 4	Compare the desirability of the feasible combinations and group those that are judged to be indifferent in terms of the criterion; each group of profiles form an equally plausible performance level of the scale (if convenient, give a label to each level). Rank the plausible levels by decreasing relative attractiveness in terms of the criterion.	This requires holistic comparisons of multidimensional profiles which can involve considerable judgmental effort; this can be facilitated using a pairwise comparison procedure. Clearly identify the most and least attractive multidimensional levels.
Step 5	Make a textual description of each plausible performance level, as detailed as appropriate and as objective as possible.	If appropriate, use pictorial representations or mention benchmarks to complement and to give reality to each description.

Building an Additive Evaluation Model: Measurable Value Functions and Scaling Constants

Once the set of evaluation criteria E_i, i = 1, …, n and their appropriate performance descriptors X_i, i= 1, …, n are defined, the second day of the decision conference is devoted to building the evaluation model. The performance profile of a bid $\boldsymbol{x}$ can be written as (x1, … , x_n), where x_i is a specific performance level of X_i. As said before, the model is an additive value function model of the form:

$$v(x_1,\cdots,x_n)=\sum_{i=1}^{n} w_i\, v_i\,(x_i) \text{ with } \sum_{i=1}^{n} w_i = 1,\ w_i > 0 \text{ and } \begin{cases} v_i\,(x_i^{+}) = 100 \\ v_i\,(x_i^{0}) = 0 \end{cases} \text{for } i = 1, \ldots, n \quad (1)$$

where v *is* the overall value score of bid $\boldsymbol{x}$ that measures its global attractiveness; v_i, i = 1, …, n are single attribute value functions, x_i^{+} and x_i^{0}, i = 1, …, n are, respectively, the 'good' and 'neutral' performance levels defined for each performance descriptor X_i, i = 1, …, n, and w_i, i = 1, …, n are scaling constants (hereafter simply taken as the weights of the criteria).

The 'most economically advantageous' bid is calculated by *max v* in model (1). The use of good and neutral levels to anchor the value function model, rather than the traditionally used most and least preferred ones, allows one to find out if a bid is attractive, neutral, or unattractive. (A bid is attractive if its overall score is better than the overall score of a hypothetical bid that has 'neutral' performances in all criteria.)

The reason for the choice of a value function model (for a decision under certainty) rather than a utility model (for a decision under risk) is that bid performances ($x1,\ldots, x_n$) can be assessed by the REN engineers with a high degree of certainty for all of the evaluation criteria E_i, $i = 1, \ldots, n$ given their confidence that the information requested of the proposers in the call for tenders is precise enough to permit one specific performance level x_i of each performance descriptor X_i, $i = 1,\ldots, n$ to be easily assigned to each bid $\mathbf{x}$. Concerning additive aggregation, the appropriate axiomatic basis in the context of bid evaluation is Dyer and Sarin's (1979) theory of "measurable value functions" which requires difference independence conditions. The alternative theory of conjoint measurement that also yields an additive value function, although requiring weaker independence conditions, is not adequate in bid evaluation because one not only wants to measure the overall attractiveness of bids but also wants to score them independently in terms of their (partial) attractiveness relative to each one of the criteria separately.

2.3 Assessing the Value Function: The MACBETH Approach

We believe, as do von Winterfeldt and Edwards (1986), "that people can make interval and ratio judgments about differences in attractiveness" (p. 212) but that they can find direct rating questions "hard to answer" (p. 210). To ease this task, MACBETH tries to answer the following question, here adapted to a performance descriptor X_i of the REN case: How can a measurable value function be built on X_i, both in a qualitatively and quantitatively meaningful way (French 1986), without forcing the WG to produce direct numerical representations of attractiveness and involving only two levels of X_i for each judgment required from the WG? Suppose, for simplicity, that the m performance levels of a discrete performance descriptor X_i are ranked in order of decreasing attractiveness, from the most preferred to the least preferred level, such that level x_{ip} is at least as attractive as level x_{ir} for $p < r$, with $p = 1, ..., m\text{-}1$ and $r = 2, ..., m$. The range of a continuous performance descriptor can be split into a few intervals and the m limits of these intervals (not the intervals) can be taken as reference levels that verify a similar ordinal assumption. If X_i is discrete, each point on the respective value function v_i will be assessed, if X_i is continuous, the points corresponding to the reference levels will be assessed and a piecewise linear approximation

of v_i will be used. Whatever the type of performance descriptor, the MACBETH questioning procedure consists of asking the WG for a qualitative judgment about the perceived difference in attractiveness between pairs of levels x_{ip} and x_{ir}. If no difference is felt, x_{ip} and x_{ir} are indifferent and therefore $v_i(x_{ip})–v_i(x_{ir}) = 0$ (the difference is 'null'). If a difference is felt, x_{ip} is strictly preferred to x_{ir} and therefore $v_i(x_{ip})–v_i(x_{ir}) > 0$. The WG is then invited to judge, qualitatively, how large the difference is, expressed in terms of a set of six semantic categories (C_k, k = 1, …, 6) of difference in attractiveness, C_1: 'very weak' (or 'between null to weak'), C_2: 'weak', C_3: 'moderate' (or 'between weak to strong'), C_4: 'strong', C_5: 'very strong' (or 'between strong to extreme') and C_6: 'extreme'. Bana e Costa and Vansnick (1994) initially called the 'weak', 'strong' and 'extreme' categories the fundamental ones, although the M-MACBETH software that implements the MACBETH approach does not make this distinction and even allows for group judgments that do not distinguish between several consecutive categories, such as 'strong or very strong'. This type of judgmental procedure motivates discussion and learning within the WG, contributing to the development of a group value system.

The group judgments are used to populate the upper triangular part of a 'MACBETH matrix of judgments' (see example in Figure 4). As each judgment is entered into the matrix, its consistency with the judgments already inserted into the matrix is checked and if an inconsistency is detected, suggestions to overcome it are presented to the WG. Technically, this is done by a mathematical programming algorithm (see Bana e Costa et al. 2005a for details).

The WG can make a number of pairwise comparisons ranging from a maximum of $m(m\text{-}1)/2$ judgments, when all pairwise comparisons are made, to a minimum acceptable number of $m–1$ judgments, as when comparing only each two consecutive levels or one level with all of the other $m–1$. However, it is recommended to ask for some additional judgments to perform "a number of consistency checks" (von Winterfeldt and Edwards 1986, p. 228). Bana e Costa and Chagas (2004) recommend filling in the border of the upper triangular portion of the matrix, for a total of $3(m\text{-}2)$ pairwise comparisons; on the other hand, Belton and Stewart 2002 (p. 173) present an example in which only $2m\text{-}3$ judgments are made (between each two levels x_{ip} and x_{ir} for $r = p + 1$ and $r = p + 2$). They also observe that (p. 173) although different evaluators may interpret the semantic labels differently, it will be true that if the difference between x_{ip} and x_{ir} is assigned to a higher attractiveness category than the difference between $x_{ip'}$ and $x_{ir'}$, then $[v_i(x_{ip})–v_i(x_{ir})] – [v_i(x_{ip'})–v_i(x_{ir'})] > 0$. These inequalities are strict because a cornerstone of the MACBETH approach is that all of the

differences allocated to one semantic preference difference category are strictly larger than those allocated to a lower category.

More specifically, each qualitative category C_k, $k = 1, \ldots, 6$ is quantitatively represented by an interval of positive real numbers, delimited in MACBETH by thresholds s_k, $k = 1, \ldots, 6$ such that $s_{k+1}–s_k \geq 1$, $k = 1, \ldots, 5$ (the minimal length of any category C_k, $k = 1, \ldots, 5$, is therefore equal to 1). This enables the strict inequality $[v_i(x_{ip})–v_i(x_{ir})] – [v_i(x_{ip'})–v_i(x_{ir'})] > 0$ to be translated by: $s_k + ½ \leq v_i(x_{ip}) – v_i(x_{ir}) \leq s_{k+1} – ½$ if the difference in attractiveness between x_{ip} and x_{ir} is assigned to category C_k, $k = 1, \ldots, 5$; and $s_6 + ½ \leq v_i(x_{ip}) – v_i(x_{ir})$ if the difference in attractiveness between x_{ip} and x_{ir} is assigned to category C_6. These new inequalities imply that $[v_i(x_{ip}) – v_i(x_{ir})] – [v_i(x_{ip'}) – v_i(x_{ir'})] \geq 1$ if the differences in attractiveness between x_{ip} and x_{ir} and between $x_{ip'}$ and $x_{ir'}$ are assigned to two different categories, except if the difference between x_{ip} and x_{ir} is assigned to 'very weak' and $x_{ip'}$ and $x_{ir'}$ are indifferent ('no' difference in attractiveness) which implies $s_1 = ½$.

For a consistent matrix of MACBETH judgments, the MACBETH scale is obtained by a linear program that minimizes $v_i(x_{i1})$ subject to the above defined constraints and the constraint $v_i(x_{ip})–v_i(x_{ir}) = 0$ if x_{ip} and x_{ir} are indifferent (plus the usual constraints of non-negative variables). Both the objective function and the set of constraints are simple enough to allow for determining the MACBETH scale by a straightforward 'hand procedure.' (For further technical details see Bana e Costa et al. 2005a.) Simple examples can easily be presented to the evaluators, thus avoiding the 'black box' effect of linear programming (see Bana e Costa 2007). The advantage of using software to do the calculations is that it can determine the interval within which each score, $v_i(x_{ip})$ for $p = 1, \ldots, m$, can vary when the other m-1 scores are fixed and still remain consistent with the pairwise preference difference assessments. This allows the adjustment of the scale by the WG while remaining consistent with the qualitative preference difference assessments. However, in applications, evaluators have often preferred to revise some of their qualitative judgments. As in any other assessment procedure, "the scale construction process stops when the decision maker is comfortable with the assessments" (von Winterfeldt and Edwards 1986, p. 228).

3. Implementing the Process: The Kick-Off Meeting

In the kick-off meeting with the WG, Henrique Gomes explained the reasons for developing a new evaluation system and invited the participants to openly discuss the pros and cons of the current bid evaluation system. From this meeting the following issues arose:

1. Although the WG agreed that invitations to tender should continue to be sent only to pre-qualified suppliers, they determined that the process by which suppliers were pre-

qualified should be revised. However, it was agreed that this issue be handled separately from the revision of the bid evaluation system. (This is, therefore, out of the scope of this paper – for a survey on contractor pre-qualification models see El-Sawalhi et al. 2007.)

2. Since the criteria relevant to the evaluation of bids, as well as their respective 'importance', vary across services, REN had been tailoring an evaluation model to each type of invitation to tender. An analysis of the criteria used across the various models revealed that some were ill-defined, others redundant, and that some relevant concerns were missing entirely from the evaluation. The WG decided that REN would continue to use separate evaluation models for different types of services, and that the number of models would be determined through the discussion and structuring of similar evaluation criteria.
3. Given the amount of money involved and the pressure applied by bidders, bid evaluations should be as transparent and objective as possible. The existing REN methodology fell short of this goal since it required REN engineers to analyze bids by directly assigning numerical scores to each bid on each criterion (except Cost). This method forced the engineers to combine factual information with value judgments, making it impossible to discern whether a given score reflected the bid's level of performance on the criterion or the attractiveness of the performance of the bid on the criterion (or both). The WG decided that REN engineers, who are technical experts, should only be asked to assess the bids' performance on the criteria. This required that the direct scoring of bids be abandoned and a new approach that addressed these issues be developed. To do so, the WG started by developing a descriptor of performance on each criterion.
4. The WG understood that, once defined, the criteria performance descriptors would serve as the operational basis for evaluating the performance of the bids received. It was important for them to note that there is a distinction between performance of a bid and its value (or attractiveness) to REN, and that the latter depends on the objectives of the call for tenders. In addition, as explained in Section 2.1, the REN evaluation model should be reusable and our proposal of building value functions to systematically 'transform' bid performances into value scores measuring their attractiveness was explained (Sections 5.1 and 5.2).
5. Directive 2004/18/EC establishes that: "Where the contracting authorities choose to award a contract to the most economically advantageous tender, they shall assess the tenders in order to determine which one offers the best value for money. In order to do

this, they shall determine the economic and quality criteria which, taken as a whole, must make it possible to determine the most economically advantageous tender for the contracting authority" (paragraph 6, page L 134/121). The WG understood that the Directive would be respected if the values of a bid across multiple criteria were aggregated through a compensatory procedure, such as an additive value model. However, assessing the 'value for money' required that 'value' be separated from 'money'. Therefore, the 'cost' criterion should be separate from the 'benefit' criteria and a multicriteria cost-benefit model should be developed.

6. Ensuring that the model met additive independence conditions, both through the structuring and building phases, was necessary to ensure coherence. However, because the former bid evaluation model used at REN included criteria that were not mutually preferentially independent, interdependent aspects were, therefore, clustered into new criteria (see Section 4).
7. The WG agreed that the requisiteness of the model would be tested throughout its live development (see Section 6). Once confirmed, the model would be ready for repeated use to calculate the overall score of each bid presented in future calls for tenders, thus ensuring decision-making transparency and the equal treatment of all bids.
8. The process by which REN weighted criteria was discussed; REN had previously assigned weights by directly comparing the "relative importance" of the criteria without considering the ranges of the performance descriptor scales. Therefore, it was not surprising that in some evaluation processes a bid (b_1) with a significantly lower cost than a second bid (b_2) would often lose in the bidding process, even though b_2 only outperformed b_1 in 'low importance' criteria of technical quality. This indicated that the trade-off between cost and quality was not considering the ranges of the performance descriptor scales. The facilitator explained to the WG that assigning weights that reflect the 'importance' of the criteria, without considering the range of the performance descriptor scales and the importance of those ranges, is a major error that must be corrected (Keeney 1992, called it the "most common critical mistake"), and that weighting anchors had to be defined and adequate trade-off judgments had to be made (Section 5.3).
9. Last but not least, we faced the need to adopt a cost-benefit weighting procedure (see Section 5.4) that was both technically correct and also able to respect the REN Board of Directors' strategic desire to keep "the importance of cost equal to the importance of quality." This was because, as the Director who chaired the WG pointed out, "we now know that weights reflect the relative importance of performance ranges and not the

relative importance of the criteria, but the bidders do not necessarily know this, thus a weight of less than 50% for cost could lead them to increase their proposed prices." Obviously, the swings from pre-defined neutral to good levels in two criteria may not have the same importance for REN. So, taking advantage of the quantitative nature of the cost performance descriptor, the reference level 'good cost' would be only fixed *a posteriori* based on a cost-benefit indifference trade-off judgment (see Section 5.4).

4. Structuring the Model

The objective of evaluating bids is to identify the one that offers the best trade-off between cost and benefit. In the REN context, the Benefit criteria were not common to all types of calls for tenders. Five different models were used to evaluate the possible types of calls for tenders. The model used to evaluate bids for construction of new power lines was rebuilt first, during a two-day decision conference; this process will be described hereafter. The model for 'supply of equipment' was subsequently developed in a one-day decision conference. The decision analysis know-how transferred to the WG during this learning process made them capable of constructing the remaining models without the need of further external facilitation.

Model-structuring began with a facilitated group analysis of the criteria (and sub-criteria) used by REN to evaluate bids for construction of new power lines, to determine which should be kept in the new model. Criteria used to pre-qualify potential bidders, criteria used for screening bids, and criteria used to evaluate bids were discussed. One such discussion involved a supplier evaluation index developed by REN that assigned performance scores to suppliers based on the perceived quality of services provided in the past; this supplier rating had previously been included in the evaluation of bids. After some discussion the WG concluded that this was redundant since the index had already been used to pre-qualify suppliers. Criteria related to bidders rather than their bids and criteria relevant to screening but not to evaluating bids were similarly eliminated. The new set of evaluation criteria, arrived at through the discussion, was then analyzed to ensure exhaustiveness, leading to the definition of two new criteria and the refinement of others. Finally, two criteria were merged into a single one due to their mutual preference dependency ('methodology and critical components for deployment'). Time was not an evaluation concern because REN always imposes a compulsory deadline. The new set of seven evaluation criteria agreed upon by the WG is much more concise than the list of 19 criteria previously used by REN. The cost-benefit tree shown in Figure 2 and the six benefit criteria were defined in the decision

conference as follows (later, they were more extensively defined in new regulations (REN 2007)):

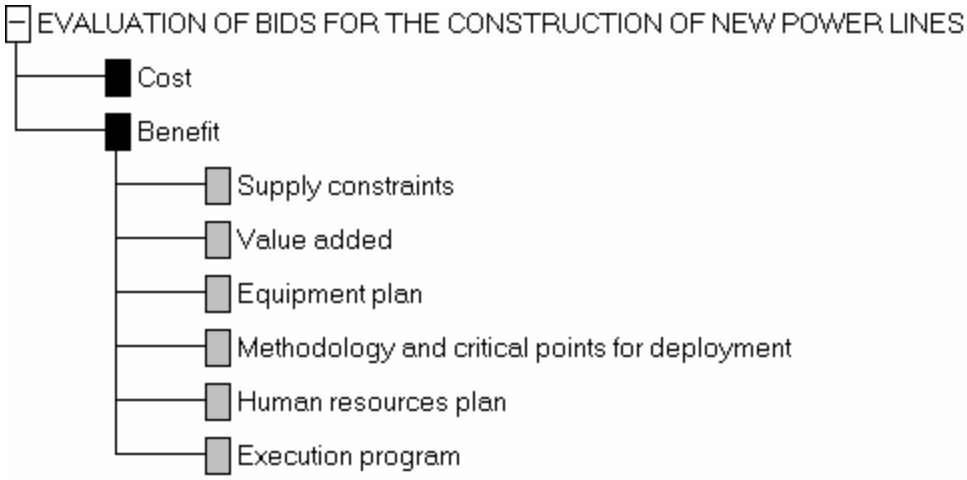


Figure 2. Evaluation criteria of bids for the construction of new power lines.

- 'Supply constraints' – the extent to which technical, logistic, administrative and financial aspects included in the bid may constrain or compromise the ability to execute in accordance with the technical specifications.
- 'Value added' – the extent to which the bid includes technical, logistic, administrative and financial aspects that add value to the supply (for instance, that add value to the most relevant equipment).
- 'Equipment plan' – the extent to which the proposed plan to allocate materials, machinery and other equipment is adequately detailed, with regard to time, types of equipment and tasks.
- 'Methodology and critical components for deployment' – the extent to which the methodology proposed is adequately explained, so as to ensure that technical specifications are respected, its components are detailed in the bid and the most critical execution issues and respective solutions are identified.
- 'Human resources plan' – the extent to which the proposed plan to allocate human resources is adequately detailed, in terms of time, and organized by functional categories and execution activities.
- 'Execution program' – the extent to which the proposed execution program is adequately detailed and organized by activities and sub-activities.

The performance of any bid on the 'cost' criterion was defined as the NPV (net present value) of the price, in euros, associated with the bid. The 'neutral cost' was set as the average of the final construction costs of a set of previous similar (in type and size) new power lines; the process to determine the 'good cost' is discussed in Section 5.4.

A constructed performance scale was developed for each one of the benefit criteria, according to the following steps: first, two reference levels, good and neutral, were defined; then, more levels were added to cover the plausible range of performances; finally, each level of the performance descriptor was carefully described to ensure a clear and unambiguous interpretation of its meaning.

Consider the 'equipment plan' criterion. The WG was first asked to define a neutral level for the criterion, i.e., define a performance on this criterion that would neither be attractive nor unattractive: 'an equipment plan which presents an allocation by type of equipment, but no time allocation'. The WG was then asked to define a good level, i.e., a performance on this criterion that would be substantially attractive: 'an equipment plan that describes the allocation of the equipment sorted by type, including when this equipment would be used'. The answers provided revealed two interrelated concerns: 'allocation by type of equipment' and 'time allocation'. The WG was asked whether any other aspects could differentiate bids (with respect only to the equipment plan). The group concluded that the plan could also offer allocations by task, adding a third characteristic to the attractiveness of an equipment plan. These are, therefore, the three characteristics that should be considered when creating an adequate description of the plausible performances of the bids with regard to their equipment plans: the bid either presents an equipment plan with or without an allocation by type of equipment, with or without time allocation, and with or without allocations by task. Figure 3 depicts a tree drawn by the facilitator with all the possible combinations of these key characteristics. Note that the WG considered some of the combinations (crossed out in Figure 3) to be unrealistic and, therefore, dropped them. The remaining combinations, marked C1 through C5, were then sorted in decreasing order of attractiveness. The WG, obviously, chose C1 to be the most attractive level of performance, since it corresponded to an equipment plan that contained all of the key characteristics (equipment/time/task) and C5 to be the least attractive level, because it corresponded to a bid that did not present an equipment plan. Both C2 and C3 presented two attractive characteristics; however, C2 was preferred to C3 because the WG deemed that having information about time was more attractive than having information about task. Finally, C3 was preferred to C4 because C4 only contained information about equipment while the former had information about equipment and task. The final step was to carefully describe each level to ensure a clear and unambiguous interpretation of its meaning. The final constructed scale is shown in Table 2.

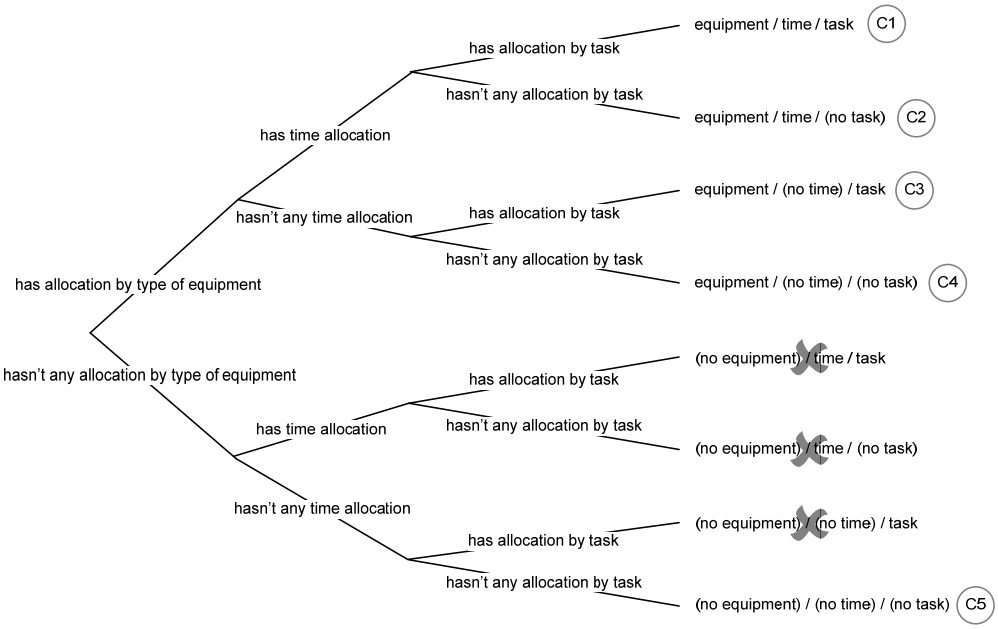


Figure 3. Possible combinations of the key characteristics of an equipment plan.

The performance descriptors for the remaining five benefit criteria were constructed through a similar process; they are shown in Tables 3 through 7. All levels of performance within each scale are presented in decreasing order of attractiveness and many of them required more detailed descriptions (that are not included in the tables) to avoid misinterpretations, either by the various REN engineers or by the same engineer at different times, when evaluating bids (see Sections 6 and 7.1).

Table 2. Constructed performance scale for the 'equipment plan' criterion.

Performance levels	
Time allocation by type of equipment and task	L1
Time allocation by type of equipment	L2=Good
Allocation by type of equipment and task (without time allocation)	L3
Allocation by type of equipment only (without time allocation)	L4=Neutral
No allocation	L5

Table 3. Constructed scale for supply constraints.

Performance levels	
The bid…	
…does not present any constraints	L1=Good
…presents one constraint that does not compromise the ability to execute according to Specifications	L2=Neutral
…presents constraints that compromise the ability to execute according to specifications	L3
…presents constraints that compromise the ability to execute according to specifications and additional constraints that do not compromise the execution according to specifications	L4

Table 4. Constructed scale for value added.

Performance levels	
Value added to relevant equipment and other value added	L1
Value added to relevant equipment	L2=Good
No value added to relevant equipment but other value added	L3
No value added	L4=Neutral

Table 5. Constructed scale for methodology and critical components for deployment.

Performance levels	
Detailed methodology; identifies critical issues and proposes solutions	L1
Generic methodology; identifies critical issues and proposes solutions	L2=Good
Detailed methodology; identifies critical issues but does not propose solutions	L3
Generic methodology; identifies critical issues but does not propose solutions	L4
Detailed methodology; does not identify critical issues	L5=Neutral
Generic methodology; does not identify critical issues	L6
The methodology is not explained	L7

Table 6. Constructed scale for human resources plan.

Performance levels	
Time allocation by category and activity	L1
Time allocation by category	L2=Good
Allocation by category and activity (without time allocation)	L3
Allocation by category only (without time allocation)	L4
Allocation by activity only (without time allocation)	L5=Neutral
It does not present any allocation	L6

Table 7. Constructed scale for execution program.

Performance levels	
Detailed by activity and sub-activity	L1=Good
Detailed by activity	L2
Generic	L3=Neutral
No execution program	L4

5. Constructing the Value Function Model with MACBETH

5.1 Building the Benefit Value Functions

A value function allows one to assign value scores to the levels of a performance descriptor relative to the fixed scores of 0 and 100 assigned to the neutral and good reference levels in the additive model (1). Therefore, a bid that outperforms (underperforms) the neutral level on a criterion will obtain a positive (negative) value score and an outstanding bid that outperforms the good level on a criterion will obtain a value score that exceeds 100 value units. In the second day of the decision conference, the MACBETH approach and software were used to support this task. As already emphasized in Section 2, MACBETH only requires qualitative judgments to generate value scales, enabling us to defocus the WG from the pitfalls of the previous REN point system and overcoming any skepticism associated with expressing judgments numerically rather than qualitatively (Gurmankin et al. 2004).

To construct a value scale for the 'equipment plan' criterion, the facilitator asked the WG to judge the differences in attractiveness between the various levels of its performance descriptor (see Table 2). The WG first judged the difference between the most preferred level L1 and the least preferred level L5. About half the members of the WG deemed this difference to be very strong, the others judged it to be extreme, so a group judgment was entered into the matrix of judgments as 'very strong or extreme'. The WG was next asked about the difference between the second most preferred level L2 and the least preferred level

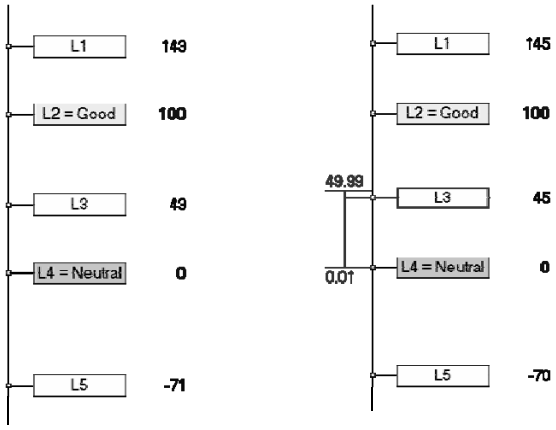
L5, which was unanimously considered to be very strong. The process continued to complete the last column of the judgments matrix; the first row of the matrix was populated next, followed by the diagonal above the main diagonal and finally, the difference between L2 and L4 was judged to be moderate. The judgments given by the WG are depicted in Figure 4. An inconsistency after the last entry occurred, because the difference between L2 and L4 should be greater than the 'strong or very strong' difference between L2 and L3 to preserve the ranking of the levels. The WG was alerted to this inconsistency (see the up-arrow and the down-arrow on Figure 4) and informed that they could either increase the difference between L2 and L4 or reduce the difference between L2 and L3, to maintain consistency across judgments. The WG decided to change the judgment between L2 and L4 to very strong.

	L1	L2 = Good	L3	L4 = Neutral	L5
L1	no	moderate	strg-vstr	v. strong	vstrg-extr
L2 = Good		no	↓ strg-vstr	↑ moderate	v. strong
L3			no	moderate	v. strong
L4 = Neutral				no	v. strong
L5					no

Figure 4. MACBETH judgments matrix for the 'equipment plan' criterion.

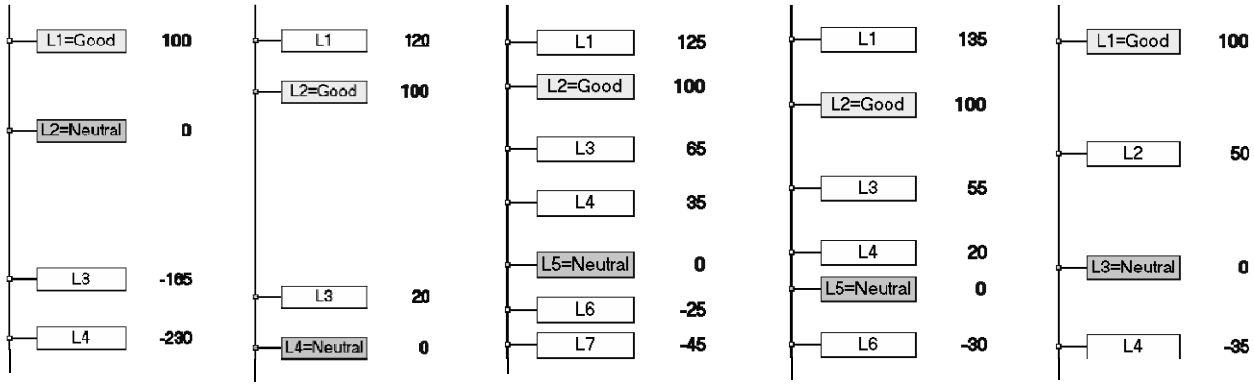
With no remaining inconsistencies, the linear programming procedure presented in Section 2.3 resulted in the scale shown in Figure 5(a), which the WG was asked to analyze, in terms of proportions of the resulting scale intervals, to ensure that their relative size correctly captured the WG's collective value judgments. The ensuing discussion led to some minor adjustments resulting in the 'equipment plan' value scale displayed in Figure 5(b). This scale allows future bid performances to be converted into value scores, e.g., a tender bid that has an equipment plan with time allocation by type of equipment and task (L1) would receive a score of 145 value units of this criterion.

Figure 6 displays charts representing the value functions obtained through a similar process for the remaining five benefit criteria.



(a) *(b)*

Figure 5. MACBETH thermometers for the 'equipment plan' criterion (a) MACBETH scale (b) interval scale.



(a) *(b)* *(c)* *(d)* *(e)*

Figure 6. MACBETH thermometers showing interval scales for the (a) 'supply constraints',(b) 'value added', (c) 'methodology and critical components for deployment', (d) 'human resources plan', and (e) 'execution program' criteria.

5.2 Defining the Cost Value Function

The bid evaluation system previously used by REN assigned a score of 100 to the lowest cost bid and a score of $100 \times \frac{\text{lowest cost}}{\text{bid cost}}$ to any other bid, which is a "normalization procedure" that "preserves proportionality", that is, using the words of (Barba-Romero 2001): "The evaluations before and after normalization exhibit the same proportions" (p. 130). The analysis of this (unfortunately popular) method spurred a long discussion regarding the preference structure implied by this expression. First, the facilitator highlighted the non-linearity of the implied value function, that is, a constant difference of cost would not have the same value difference along the cost scale. To verify the accuracy of this implicit assumption the facilitator asked the WG to provide MACBETH judgments for several equally spaced costs. The WG's answers revealed that its preference function for cost was actually linear, showing that the previous scoring procedure was inaccurate, and it was, therefore, abandoned. Moreover, the pitfall of defining a scoring system that is dependent upon the performance of a particular bid was explained to the WG, as well as how this is

incompatible with an additive value model, which requires fixed references to assess meaningful weights ('good' and 'neutral', in the REN case). The score of a given bid on Cost was then set equal to

$$100 \times \frac{\text{neutral cost - bid cost}}{\text{neutral cost - good cost}}.$$

5.3 Weighting Benefit Criteria

This section explains the process by which the benefit criteria were weighted. Section 5.4 describes how cost and benefit were assigned equal weights, in accordance with the strategic decision of the REN Board of Directors justified in Section 3.

To weight the benefit criteria, the WG started by ranking the 'good – neutral' swings by their overall attractiveness. The final ranking of the swings is shown in Figure 7. (Note that the swings in supply constraints and equipment plan were considered to be equally attractive; they were therefore assigned equal weights.) The WG, then, qualitatively judged the overall attractiveness of each swing (see the judgments in the last column of the matrix in Figure 8). Next, the most important swing was qualitatively compared to each of the others, followed by the comparison of each consecutive pairing of swings. (See the example in Figure 9.)

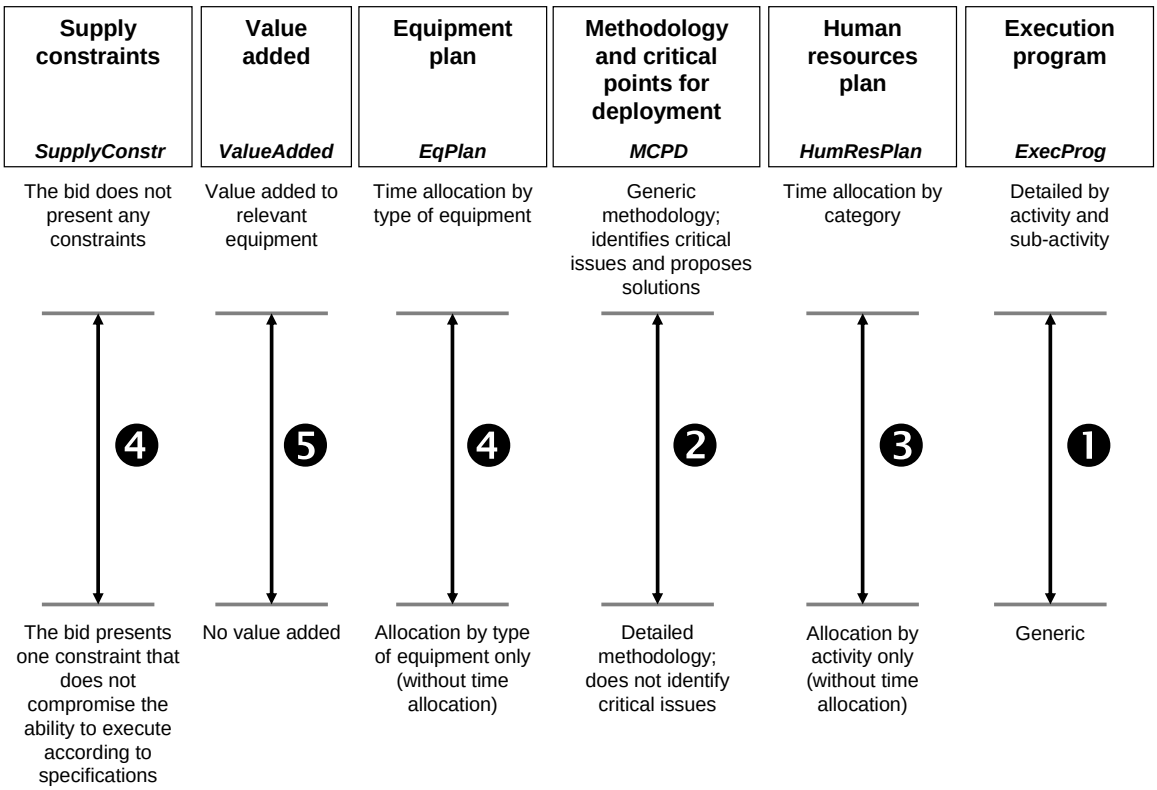
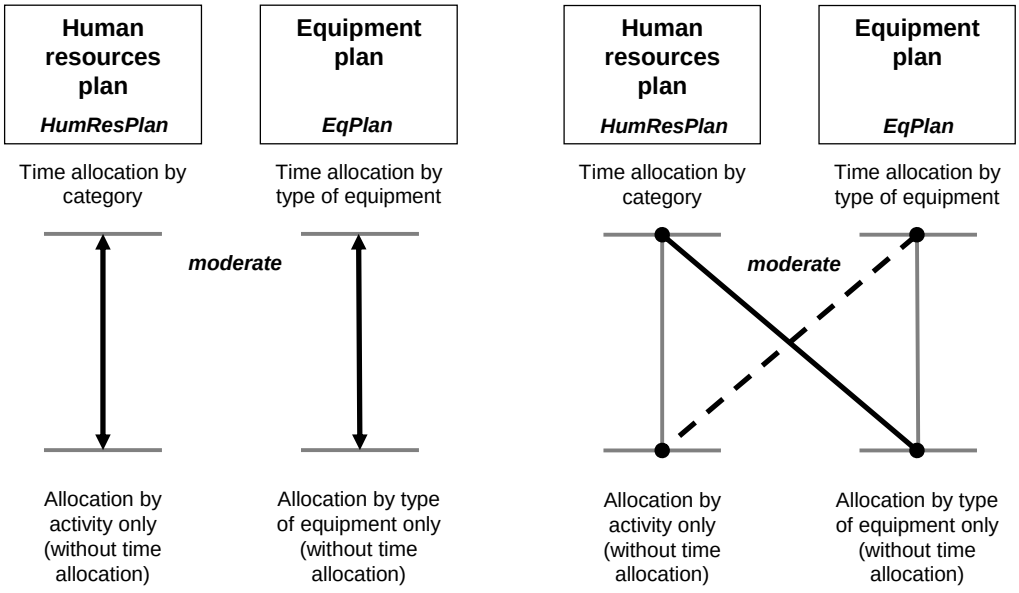


Figure 7. References 'good' and 'neutral' of the benefit criteria and ranked swings.

	[ExecProg]	[MCPD]	[HumResPlan]	[EqPlan]	[SupplyConstr]	[ValueAdded]	Neutral all over
[ExecProg]	I	weak	weak	strong	strong	v. strong	vstrg-extr
[MCPD]		I	weak	P	P	P	v. strong
[HumResPlan]			I	moderate	P	P	v. strong
[EqPlan]				I	I	P	strong
[SupplyConstr]				I	I	moderate	strong
[ValueAdded]						I	moderate
Neutral all over							I

Figure 8. Weighting judgment matrix for the benefit criteria. The P and I within the matrix respectively mean Positive difference of attractiveness and Indifference (i.e., no difference of attractiveness).

Since the WG was satisfied with the process and did not wish to provide additional judgments, the MACBETH scale of swing weights was presented, discussed and adjusted. The final relative weights for the benefit criteria are presented in Figure 10.



(a) *(b)*

Figure 9. Difference of attractiveness between the swings of 'human resources plan' and 'equipment plan'.

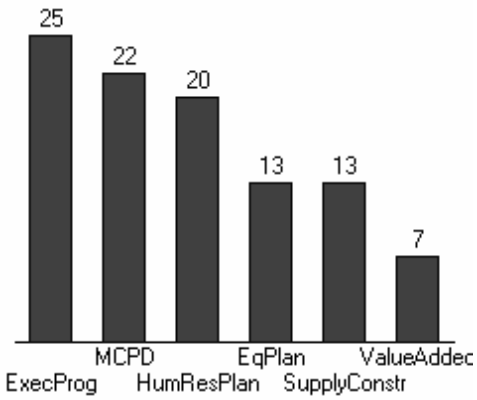


Figure 10. Weights of the benefit criteria.

It should be noted that the previous weighting steps included judgments that required the acceptance of compensations between performances in two different criteria. The fact

that the WG has expressed such judgments validates, from a constructive perspective of decision aiding, the working premise of additive aggregation. However, note also that, in practice, the validation (or not) of this fundamental premise of compensation is both substantive and theoretically meaningless if weighting references were not previously defined on the criteria. For this reason, it is prudent for evaluation systems not to allow users to directly assign weights to criteria (see Liu et al. 2000, p. 368, or Barba-Romero 2001, p. 127). Finally, the legal requirement that weights should be announced in the call for tenders, that is, before the bids are presented, does not require the use of direct weighting procedures, as explained in Bana e Costa et al. (2002).

5.4 Cost-Benefit Trade-Off

As explained in Section 3, the REN Board determined that the weight of the cost criterion was to equal that of the benefit criterion, i.e., the weight of cost had to equal 0.5, as did the sum of the weights of all benefit criteria. (The relative weights of the benefit criteria shown in Figure 10 should, therefore, be divided by 2.) Requiring cost and overall benefit to be equally weighted meant that a reduction in cost from 'neutral cost' to 'good cost' would have to be as attractive as an increase from a 'neutral benefit' to a 'good benefit'. Note that three of these four references had already been fixed. As mentioned in Section 4, the 'neutral cost' was fixed as the average of the real costs of previous, similar (in type and size), projects. Good and neutral benefits were fixed implicitly, a bid with neutral performances on all benefit criteria (as defined in Tables 2 through 7) was defined to provide a neutral overall benefit; similarly, a bid with good performances on all benefit criteria was defined to provide a good overall benefit. Consequently the only reference point that was not yet fixed was the 'good cost', which had to be defined in such a way that a reduction in cost from 'neutral cost' to 'good cost' perfectly compensated a decrease from 'good benefit' to 'neutral benefit'. The WG easily concluded that the trade-off between cost and benefit should not be the same for projects with significantly different neutral costs. Therefore, two scenarios were considered. In the first scenario the following two hypothetical bids were defined (see Figure 11):

B - **Costs € 1.2 million** and provides a **good benefit**.

C - **Costs € *x*** and provides a **neutral benefit**.

The WG was then asked to define *x* in such a way that bid C would be as attractive as bid B. After some discussion, the WG agreed that the cost of C should be about 5% less than the cost of B, therefore, *x* (the 'good' cost) should equal € 1.14 million (€ 60 thousand less than neutral).

A second scenario was then defined for a smaller neutral cost:

B' - **Costs € 800 thousand** and provides a **good benefit**.

C' - **Costs € *y*** and provides a **neutral benefit**.

This time the WG fixed the trade-off at about 10% of the cost of B', consequently the 'good' cost was fixed at € 720 thousand (now, € 80 thousand less than neutral).

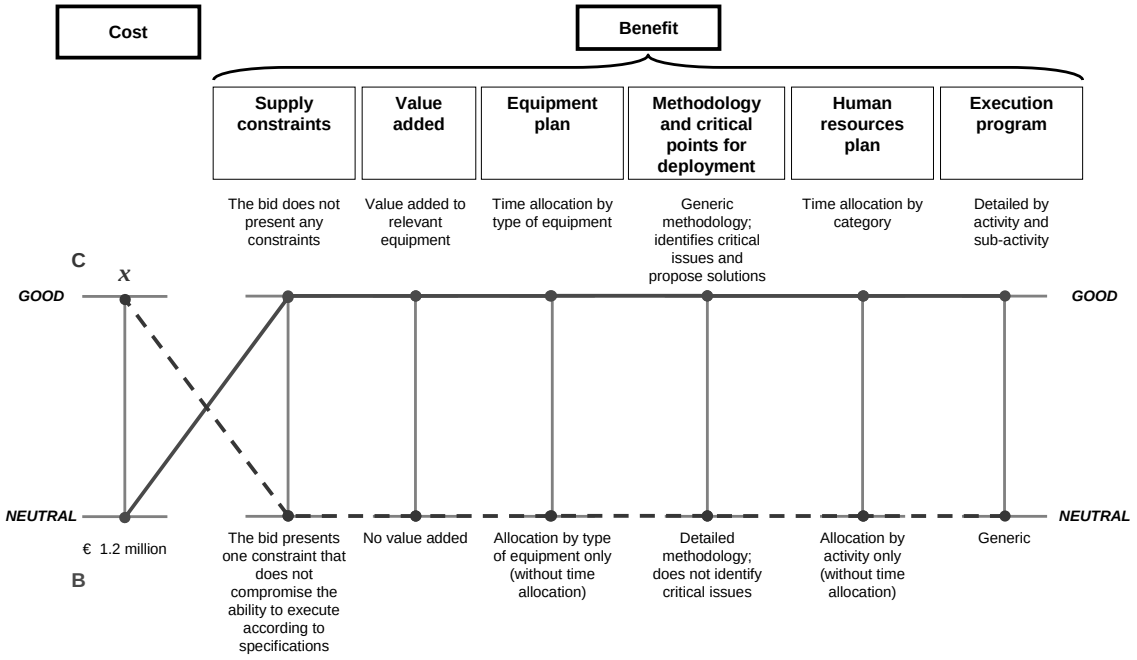


Figure 11. Trade-off between Cost and Benefit: hypothetical bid B is represented by the solid line; hypothetical bid C is represented by the dashed line.

It is interesting to note that the WG always expressed their trade-off judgments as a percentage of the neutral cost and that these percentages varied for different reference costs. This presented an interesting dilemma, how should REN determine the appropriate percentages to define the 'good' costs given the varying magnitudes of the calls for tenders? After a long discussion, the WG agreed upon the following rules: (i) for a neutral cost lower than € 0.9 million, the 'good' cost should be 10% less than the neutral cost; (ii) for a neutral cost greater than € 1.1 million, the 'good' cost should be 5% less than the neutral cost; (iii) for a neutral cost between € 0.9 million and € 1.1 million, the 'good' cost should be *p*% less than the neutral cost, with *p* decreasing linearly between 10 and 5.

6. Validating the 'Requisiteness' of the Model

Was the new model a requisite model (Phillips 2007); that is, was it adequate to evaluate the bids while respecting the REN value system? If not, the model should have been fine-tuned or, if necessary, rebuilt. A good way to have tested the model would have been to use historical REN bid data. Unfortunately, much of the detailed data necessary to do so did not exist. As a result, several scenarios were built and M-MACBETH was used to test the model

using these scenarios. Hypothetical bid performances were entered into the software, but, for some Benefit criteria the WG encountered difficulties in deciding which level of performance to select. For instance, the WG considered that the performances of some bids in terms of equipment plan did not correspond to any performance level of the performance descriptor in Table 2, but rather to some intermediate levels. For example, a bid may exist that contains an equipment plan with time allocation by type of equipment but ignores some, or even the majority, of the tasks to be performed by the equipment. For these types of circumstances intermediate performance levels (such as L2++, L2+ and L4/L5) were added to the constructed scale, allowing for finer distinctions in bid performance appraisal (see Table 8). The scores of the new intermediate levels were determined by linear interpolation (see Figure 12) and were validated by the WG. To address similar issues in other Benefit criteria, several intermediate performance levels were also added to their performance descriptors (see Section 7.1).

Table 8. Updated constructed scale for 'equipment plan'.

Performance levels	
Time allocation by type of equipment and task	L1
L2++	L2++
L2+	L2+
Time allocation by type of equipment	L2=Good
Allocation by type of equipment and task (without time allocation)	L3
Allocation by type of equipment only (without time allocation)	L4=Neutral
L4/L5	L4/L5
No allocation	L5

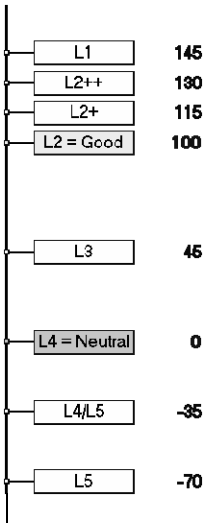


Figure 12. The final MACBETH thermometer for 'equipment plan'.

7. Discussion and Conclusion

7.1 Model Usage Feedback

On October 2007, about two years after our intervention at REN, a meeting with the managers responsible for bid evaluation took place to get their feedback about (1) the extent

to which the new models overcame the shortcomings presented by the legacy models, (2) whether the new models are still well adjusted to the requirements of REN, (3) how comfortable the WG participants had felt when they constructed other evaluation models without our assistance, and (4) how comfortable the REN analysts are with the models' usage. The meeting was held at the REN headquarters and included three of the six members of the WG, namely Jorge Liça (Head of the Equipment Division), Luís Brito da Cruz (Construction Manager) and Alberto Costa (Quality, Environment and Safety Manager). They informed us that the models developed with the MACBETH approach have been intensively used and said that: "During the first 22 months of usage we evaluated the bids submitted to 127 calls for tenders, awarding contracts summing up to around €350 million." In this period the REN analysts felt no need to further adjust the models, be it on the criteria, descriptors of performance, weights or value functions. In addition to the outcome, REN also valued the process of constructing the models: "We found this to be the right approach since it enabled us to understand all of the components of the problem. Also, it involved all group members in all phases, which permitted us to understand certain aspects thereto not understood by all members of the group. Finally, it allowed the commitment to the model to be a product of the group." Furthermore, the models were so well aligned with the REN's value system that they were able to develop new models for other purposes by themselves. When questioned about the ease of use experienced by the engineers in charge of the appraisal of bid performances, they stated: "The engineers who analyze bids have no problems using the models developed, namely because they were coached in training sessions before using the models. The criteria and their descriptors of performance were discussed in an effort to ensure unambiguous understanding of their meaning, so that a specific bid would be appraised similarly on each criterion regardless of the person making that appraisal." Table 9 contains the final performance descriptor for 'supply constraints', composed of the four initial performance levels (described in Table 3) and expanded by four intermediate performance levels added during the validation of the requisiteness of the model (addressed in Section 6). The intended meaning of each level of the descriptor is clarified, if necessary, by associating to it relevant examples of real world bid performances (see second column of Table 9; note that no examples of intermediate level L3/L4 have been encountered, which suggests that this level might eventually be dropped). The larger the number of examples available for the levels of a constructed scale, the more clearly they can be defined; this is particularly important for intermediate levels, which tend to be roughly defined. Records of all bid evaluations performed in the last two years include many relevant examples; however, these are kept confidential by REN.

Table 9. Expanded constructed scale for supply constraints.

Performance levels	Examples of corresponding bid performances
L1=Good: The bid does not present any constraints	–
L2=Neutral: The bid presents one constraint that does not compromise the ability to execute according to specifications	←The bid states that REN must guarantee that agreements have been reached with the owners of all the properties in which pylons will be installed, to ensure compliance with the deadlines outlined in the work plan. ←The bid states that the prices of either pylons or cables are revisable.
L2-	←The bid states that both the prices of pylons and cables are revisable. ←The bid states that the equipment is in the process of being purchased.
L2--	←The bid states that the prices of either pylons or cables are revisable; the bid also states that REN must guarantee that agreements have been reached with the owners of the properties in which all pylons will be installed, to ensure compliance with the deadlines outlined in the work plan; the bid also states that tracking pylon legs used on uneven surfaces is of value.
L2---	←The bid shows that the supplier is unable to meet the time limit set by REN for deactivating the power line to which the new line must connect.
L3: The bid presents constraints that compromise the ability to execute according to specifications	←The bid states that the power line to which the new line will connect must be deactivated during the total duration of the project. ←The bid states that the supplier is unable to meet the time limit set by REN for deactivating the power line to which the new line will connect; the bid also states that REN must guarantee that authorization has been granted to install at least 50% of the pylons, to ensure compliance with the deadlines outlined in the work plan.
L3/L4	(Real world examples have not yet been encountered)
L4: The bid presents constraints that compromise the ability to execute according to specifications and additional constraints that do not compromise the execution according to specifications	←The bid states that the prices of either pylons or cables are revisable; the bid also states that the power line to which the new line will connect must be deactivated during the total duration of the project.

At the beginning of 2006, around 30 evaluation processes took place during a two month period. At this time the newly developed models were tested in parallel with the REN legacy models. According to the REN managers interviewed, "the new models outperformed the previous models. With the previous models, we were sometimes obliged to fix problems caused by the unbalance between the cost and quality criteria; with the new models this is no longer necessary as these problems have ceased to exist." Furthermore: "The new models have been in continuous use for two years and we still consider them to be good. Not only do they overcome the shortcomings identified in the previous models but also the results of the evaluations have not been challenged by the bidders. Every time a bidder wants some clarification regarding the reason it did not win a particular bid evaluation process, we show

them the evaluation details, which, so far, have led to no further complaints. The same did not happen with the evaluations made using the previous models, which experienced some, albeit informal, complaints." Another valued feature of the new models is that "they allow us to transform qualitative evaluations of the bids into quantitative scores." In fact, the evaluators no longer have to directly assign numerical scores to each bid; instead they only have to choose the qualitative level that best reflects the performance of the bid on each benefit criterion. Finally, when queried about the MACBETH questioning procedure, the REN managers said that "qualitative evaluations are more intuitive." This has been confirmed by similar reactions from participants in other applications of the MACBETH approach to decision aiding in different public sector contexts (see, for example, Bana e Costa et al. 2001, 2004, 2006, 2008).

7.2 Conclusion

Based on these results, the MACBETH methodology, when used within a multicriteria decision conferencing process, not only improved the WG understanding of the key issues, but also quickly created formal bid evaluation models that were easily understood by the REN decision makers, users and suppliers. It is important to mention that the models created were publicized to REN suppliers and information about criteria and weights is currently available on-line on the REN website (www.ren.pt). This contributed to the internal and external transparency of the REN bid evaluation process and to the increased quality of bids, which have further benefited from the publication of scores obtained by suppliers in previous bid evaluations. Finally, this methodology to construct bid evaluation models is not only suited to electric transmission companies, it is valuable for any organization (public or private) that regularly evaluates procurement bids, especially when circumstances require that performance descriptors, value scales and criteria weights be defined in advance.

It is worth mentioning that some of the methodological steps followed are not exclusive to the MACBETH methodology, such as the procedures for developing and validating constructed descriptors of performance or the trade-off procedure used to construct a cost-benefit evaluation model with equal weights for 'cost' and 'overall benefit'. Instead, they can be combined with other value function assessment approaches. Doing so, however, may sacrifice the added value of using a qualitative pairwise comparison approach.

Acknowledgments

The authors gratefully acknowledge that this paper is based upon work developed under a contract between REN and IST (Instituto Superior Técnico) and would like to thank both institutions for the opportunity to publish the case and Henrique Gomes, Jorge Liça, Alberto

Costa and Luís Brito da Cruz for their agreement to allow comments they made about the process and its outputs to be reproduced in the paper. We express our gratitude to Alan White, Alec Morton, Barbara Fasolo, two anonymous referees and an associate editor for their thorough and insightful comments on earlier versions of this paper. The ideas and suggestions of Jean-Marie De Corte, Jean-Claude Vansnick and Larry Phillips significantly contributed to this paper.

References

Bana e Costa C.A., P. Antão da Silva, F.N. Correia. 2004. Multicriteria evaluation of flood control measures: The case of Ribeira do Livramento. *Water Resources Management* **18**(3) 263-283.

Bana e Costa C.A., E. Beinat. 2005. Model-structuring in public decision-aiding. Working Paper LSEOR 05.79, London School of Economics, London, UK, www.lse.ac.uk/collections/operationalResearch/research/workingPapers.htm.

Bana e Costa C.A., M.P. Chagas. 2004. A career choice problem: an example of how to use MACBETH to build a quantitative value model based on qualitative value judgments. *European Journal of Operational Research* **153**(2) 323-331.

Bana e Costa C.A., E.C. Corrêa, J.M. De Corte, J.C. Vansnick. 2002. Facilitating bid evaluation in public call for tenders: a socio-technical approach. *Omega* **30**(3) 227-242.

Bana e Costa C.A., J.M. De Corte, J.C. Vansnick. 2005a. On the mathematical foundations of MACBETH. J. Figueira, S. Greco, M. Ehrgott, eds. *Multiple Criteria Decision Analysis: State of the Art Surveys*. Springer, New York, 409-442.

Bana e Costa C.A., J.M. De Corte, J.C. Vansnick. 2005b. M-MACBETH Version 1.1 User's Guide, http://www.m-macbeth.com/downloads.html#guide.

Bana e Costa C.A., T.G: Fernandes, P.V.D. Correia. 2006. Prioritisation of public investments in social infra-structures using multicriteria value analysis and decision conferencing: A case-study. *International Transactions in Operational Research* **13**(4) 279-297.

Bana e Costa C.A., F. Nunes da Silva, J.C. Vansnick. 2001. Conflict dissolution in the public sector: A case-study. *European Journal of Operational Research* **130**(2) 388-401.

Bana e Costa C.A., C.S. Oliveira, V. Vieira. 2008. Prioritization of bridges and tunnels in earthquake risk mitigation using multicriteria decision analysis: Application to Lisbon. *OMEGA* **36**(3) 442-450 (doi: 10.1016/j.omega.2006.05.008).

Bana e Costa C.A., J.C. Vansnick. 1994. MACBETH – An interactive path towards the construction of cardinal value functions. *International Transactions in Operational Research* **1**(4) 489-500.

Bana e Costa C.A., J.C. Vansnick. 1997. Applications of the MACBETH approach in the framework of an additive aggregation model. *Journal of Multi-Criteria Decision Analysis* **6**(2) 107-114.

Bana e Costa C.A., J.C.Vansnick. 1999. The MACBETH approach: Basic ideas, software and an application. N. Meskans, M. Roubens, eds. *Advances in Decision Analysis*. Kluwer Academic Publishers, Dordrecht, 131-157.

Bana e Costa J. 2007. Behind MACBETH. Presented at the Nato Advanced Research on Risk, Uncertainty and Decision Analysis for Environmental Security and Non-chemical Stressors, Lisbon, Portugal, http://www.m-macbeth.com/references.html#basic.

Barba-Romero S. 2001. The Spanish government uses a discrete multicriteria DSS to determine data-processing acquisitions. *Interfaces* **31**(4) 123-131.

Barron F.H., H.B. Person. 1979. Assessment of multiplicative utility functions via holistic judgments. *Organizational Behavior and Human Performance* **24**(2) 147-166.

Beinat E. 1997. *Value Functions for Environmental Management*. Kluwer Academic Publishers, Dordrecht.

Belton V. 1985. The use of a simple multiple-criteria model to assist in selection from a shortlist. *The Journal of the Operational Research Society* **36**(4) 265-274.

Belton V., T.J. Stewart. 2002. *Multiple Criteria Decision Analysis: An Integrated Approach*. Kluwer Academic Publishers, Boston, MA.

Buede D.M., T.A. Bresnick. 1992. Applications of decision analysis to the military systems acquisition process. *Interfaces* **22**(6) 110-125.

Butler J.C., J.S. Dyer, J. Jia. 2006. Using attributes to predict objectives in preference models. *Decision Analysis* **3**(2) 100-116.

Dyer J.S., H.W. Lorber. 1982. The multiattribute evaluation of program-planning contractors. *Omega* **10**(6) 673-678.

Dyer J.S., R.K. Sarin. 1979. Measurable multiattribute value functions. *Operations Research* **27**(4) 810-822.

Edwards W., F.H. Barron. 1994. SMARTS and SMARTER: Improved simple methods for multiattribute utility measurement. *Organizational Behavior and Human Decision Processes* **60**(3) 306-325.

El-Sawalhi N., D. Eaton, R. Rustom. 2007. Contractor pre-qualification model: State-of-the-art. *International Journal of Project Management* **25**(5) 465-474.

Ewing Jr. P.L., W. Tarantino, G.S. Parnell. 2006. Use of decision analysis in the army base realignment and closure (BRAC) 2005 military value analysis. *Decision Analysis* **3**(1) 33-49.

French S. 1986. *Decision Theory: An Introduction to the Mathematics of Rationality*. Ellis Horwood, Chichester.

Gregory R., S. Lichtenstein, P. Slovic. 1993. Valuing environmental resources: A constructive approach. *Journal of Risk and Uncertainty* **7**(2) 177-197.

Gurmankin A.D., J. Baron, K. Armstrong. 2004. The effect of numerical statements of risk on trust and comfort with hypothetical physician risk communication. *Medical Decision Making* **24**(3) 265-271.

Hatush Z., M. Skitmore. 1997. Criteria for contractor selection. *Construction Management and Economics* **15**(1) 19-38.

Hatush Z., M. Skitmore. 1998. Contactor selection using multicriteria utility theory: An additive model. *Building and Environment* **33**(2-3) 105-115.

Holt G.D. 1998. Which contractor selection methodology? *International Journal of Project Management* **16**(3) 153-164.

Holt G.D., P.O. Olomolaiye, F.C. Harris. 1995. A review of contractor selection practice in the U.K. construction industry. *Building and Environment* **30**(4) 553-561.

Keeney R.L. 1992. *Value-Focused Thinking: A Path to Creative Decisionmaking*. Harvard University Press, Cambridge, MA.

Keeney R.L. 2007. Developing objectives and attributes. W. Edwards, R.F. Miles Jr., D. von Winterfeldt, eds. *Advances in Decision Analysis: From Foundations to Applications*. Cambridge University Press, Cambridge, 104-128.

Keeney R.L., H. Raiffa. 1976. *Decisions with Multiple Objectives: Preferences and Value Trade-offs*. Wiley, New York.

Keeney R.L., D. von Winterfeldt. 2007. Practical value models. W. Edwards, R.F. Miles Jr., D. von Winterfeldt, eds. *Advances in Decision Analysis: From Foundations to Applications*. Cambridge University Press, Cambridge, 232-252.

Kirkwood C.W. 1997. *Strategic Decision Making: Multiobjective Decision Analysis with Spreadsheets*. Duxbury Press, Belmont, CA.

Krantz D.H., R.D. Luce, P. Suppes, A. Tversky. 1971. *Foundations of Measurement, Volume 1*. Academic Press, New York.

Liu S.L., S.Y. Wang, K.K. Lai. 2000. Multiple criteria decision making models for competitive bidding. Y. Shi, M. Zeleny, eds. *New Frontiers of Decision Making for the Information Technology Era*. World Scientific, London, 349-72.

Mustafa M.A., T.C. Ryan (1990). Decision support for bid evaluation. *International Journal of Project Management* **8**(4) 230-235.

Oliveira R.C., J.C. Lourenço. 2002. A multicriteria model for assigning new orders to service suppliers. *European Journal of Operational Research* **139**(2) 390-399.

Phillips L.D. 2007. Decision conferencing. W. Edwards, R.F. Miles Jr., D. von Winterfeldt, eds. *Advances in Decision Analysis: From Foundations to Applications*. Cambridge University Press, Cambridge, 375-399.

Phillips L.D., M.C. Phillips. 1993. Facilitated work groups: Theory and practice. *Journal of the Operational Research Society* **44**(6) 533-549.

Phillips L.D., C.A. Bana e Costa. 2007. Transparent prioritisation, budgeting and resource allocation with multi-criteria decision analysis and decision conferencing. *Annals of Operations Research* **154**(1) 51-68.

Pongpeng J., J. Liston. 2003. TenSeM: A multicriteria and multidecision-makers' model in tender evaluation. *Construction Management and Economics* **21**(1) 21-30.

REN. 2007. *Processo de Concurso: Programa de Concurso (PC-001, Edição: 03, Junho/2007)*. www.ren.pt/content/CDE331E0575747F0A84E4C2519D1258.PDF

Sarin R.K., A. Sicherman, K. Nair. 1978. Evaluating proposals using decision analysis. *IEEE Transactions on Systems, Man and Cybernetics* **8**(2) 128-131.

Schein E. 1999. *Process Consultation Revisited: Building the Helping Relationship*, Addison-Wesley, Reading, MA.

von Winterfeldt D., W. Edwards. 1986. *Decision Analysis and Behavioral Research*. Cambridge University Press, New York.