



Desirability–doability group judgment framework for the collaborative multicriteria evaluation of public policies

Carlos A. Bana e Costa^{a,b}, Mónica D. Oliveira^{a,c} , Teresa C. Rodrigues^a
and Ana C.L. Vieira^{a,*} 

^aCEG-IST, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, Lisbon 1049-001, Portugal

^bLSE Health–Medical Technology Research Group (MTRG), London School of Economics, Houghton St, London WC2A 2AE, UK

^ciBB- Institute for Bioengineering and Biosciences and i4HB- Associate Laboratory Institute for Health and Bioeconomy, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, Lisbon 1049-001, Portugal

E-mail: carlosbana@tecnico.ulisboa.pt [Bana e Costa]; monica.oliveira@tecnico.ulisboa.pt [Oliveira];
teresacrodrigues@tecnico.ulisboa.pt [Rodrigues]; ana.lopes.vieira@tecnico.ulisboa.pt [Vieira]

Received 28 December 2021; received in revised form 4 November 2022; accepted 8 January 2023

Abstract

Desirability–doability framework ($2 \times D$) is a novel framework for the collaborative evaluation of public policies. Fundamental objectives and performance indicators are agreed upon in workshops, policies are characterised, and barriers to implementation identified. MACBETH interactive protocols are then applied in decision conferences to elicit qualitative judgments about the desirability of policies, within and across objectives; and about their doability under the expected graveness of barriers on contrasting scenarios. Elicited judgments allow, respectively, to construct a shared multicriteria model measuring the overall desirability of policies; and, to measure their doability. Desirability–doability graphs enable visual interactive classification of policies, with sensitivity/robustness analyses of uncertainties. $2 \times D$ was successfully tested in a real-world urban-health policymaking case to evaluate spatial policies. The main novelty of $2 \times D$ is that it bridges the socio-technical gap, present in OR, between the support required by a complex social decision-making process, and that usually offered by analytic techniques – while keeping modeling theoretically sound and simple.

Keywords: policy evaluation; socio-technical framework; desirability; doability; elicitation protocols; multicriteria analysis; MACBETH; scenarios

1. Introduction

Robust evaluation frameworks are critical for informing policymakers about the relevance of competing and complementary policies in a transparent and structured way. Clarifying policy implications and providing arguments and criteria to evaluate them confers legitimacy on the decision

*Corresponding author.

© 2023 The Authors.

International Transactions in Operational Research published by John Wiley & Sons Ltd on behalf of International Federation of Operational Research Societies

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

made (Morestin, 2012). Aiding policymakers in strengthening their capacity to make good use of information, tacit knowledge, and analytical frameworks (Head, 2016) is crucial. Applications of multicriteria decision analysis (MCDA), namely, multicriteria value measurement (von Winterfeldt and Edwards, 1986; Kirkwood, 1997), are successful in many policymaking contexts (see, e.g., Dodgson et al., 2009; Gregory et al., 2012); however, in our view, there is still a socio-technical gap between the support required by a complex social decision-making process and that usually offered by analytic techniques. This article proposes a new socio-technical framework that is able to address five key challenges that confront the facilitator of a group process. These are:

Challenge 1. In our experience, the key challenge is related to a cognitive phenomenon involving the mixed perception of desirability and doability that arises when making judgments about potential policies. As noted by Keeney (1992, p. 3), “the relative desirability of consequences is a concept based on values.” That is, judgments about the desirability (value) of a policy should not be affected by its doability—the extent to which significant barriers to policy implementation are envisaged. For example, resistance to change may influence “the evaluation of the various change alternatives” (Pardo del Val and Martínez Fuentes, 2003). Mixing the two concepts does not allow policymakers to identify the full potential of a policy. Overcoming this mixed perception requires carefully designed judgment elicitation procedures separating the two concepts. A key idea is asking policymakers to assume that there are no barriers affecting the implementation of policies when judging their desirability.

Challenge 2. As complex policy evaluation settings involve many actors, simplicity should be a *desideratum* when designing interactive group procedures for eliciting desirability and doability judgments, separately—that is, simple enough to stimulate the expression and debate of judgments. This article proposes tailor-made qualitative interactive protocols to address this challenge, in view of successful collaborative modeling.

Challenge 3. Procedures to elicit doability judgments do not explicitly consider the uncertainties emanating from the fact that barriers to implementation may be affected by non-controllable external sources. This article reports an innovative step forward by introducing scenarios of plausible futures and eliciting doability judgments separately for each of them.

Challenge 4. Facing the complexity inherent in policy evaluation processes involving multiple objectives is a challenge that can be addressed by identifying the fundamental objectives and tackling them onto a multicriteria desirability model as evaluation criteria. This is the well-known decompositional strategy of “divide and conquer” (Morera and Budescu, 1998). Raiffa (1968) describes it in simple terms: “decompose a complex problem into simpler problems, get one’s thinking straight on these simpler problems, paste these analyses together with logical glue, and come out with a program of action for the complex problem” (p. 271). Factual data and value judgments are the input used to operationalize this procedure. Assuming that judgments about the desirability of policies on an objective are not affected by their performance on other objectives, then a simple additive multicriteria desirability model can be built to assign an overall desirability score to each policy.

Challenge 5. When constructing a multicriteria desirability model, the performances of policies are assumed as sure things when expressing desirability judgments within and across objectives; however, different sources of uncertainty can affect both performance and judgments and compromise the stability of the overall desirability scores that result from the multicriteria model

(Pelissari et al., 2021). Group distrust can arise in a model due to unstable results, and the challenge for the facilitator is to follow a path that deals with uncertainty without compromising the *desideratum* of simplicity. The uncertainties that matter can be addressed through extensive sensitivity and robustness analyses of model results.

In general, any multicriteria analytical framework is appropriate for measuring desirability (Kirkwood, 1997; Hammond et al., 1998); however, a literature search did not reveal a single analytical proposal designed to address these five challenges together. By definition, they all face Challenge 4, but only Bana e Costa et al. (2014) developed a multicriteria desirability–doability model (also applied in Mateus et al., 2017) (apart from popular *ad hoc* approaches to building impact–doability or impact–effort matrices; e.g., see Baxter, 2015). They do not, however, incorporate the construction of scenarios to account for uncertainties. This is done in work developed by several authors who only model multicriteria desirability, such as in some joint applications of scenario building and multicriteria value measurement (Durbach and Stewart, 2012; Karvetski and Lambert, 2012; Goodwin and Wright, 2014).

This article develops a novel framework for the collaborative multicriteria evaluation of public policies based on judgmental information elicited from a group of policymakers about their perceptions of a policy’s desirability, on the one hand, and doability, on the other hand. Distinguishing between *desirability* and *doability* helps achieve social alignment, and trust in the decision-aiding process and the outcomes of the analysis. The novel “desirability–doability” framework (hereafter $2 \times D$) allows the five challenges to be faced in an integrated way.

In addition to being based on sound theory, practical validation should also be undertaken to legitimize a decision-aiding process. The EURO-HEALTHY project (2015–2017) (Santana et al., 2020) offered an opportunity to live test the operationalization of the most conceptual proposals of the $2 \times D$ framework in a real-world setting of urban health policymaking and policy evaluation in Lisbon, with policy outcomes spread along the several spatial policy units that comprise the Lisbon city territory. This introduces a new context-dependent challenge to be addressed, and its relevance for public policy evaluation makes it worth incorporating in $2 \times D$:

Challenge 6. How can the consequences of spatial policies be described for the easier elicitation of desirability judgments?

Following this introduction, Section 2 describes the $2 \times D$ framework for policy evaluation as developed through a group interactive four-phase socio-technical process. The Lisbon case is described in Section 3 as a proof of concept of the feasibility of implementing each part of $2 \times D$. Section 4 discusses some of the core methodological options behind $2 \times D$ and provides practical insights from *ex-post* feedback given by the Lisbon case participants. Section 5 concludes the article.

2. The multicriteria $2 \times D$ framework for policy evaluation

2.1. Overview of the $2 \times D$ framework

The $2 \times D$ socio-technical framework provides a decision-aiding guide for facilitators regarding how to conduct a sequence of four interactive group modeling phases to overcome the challenges

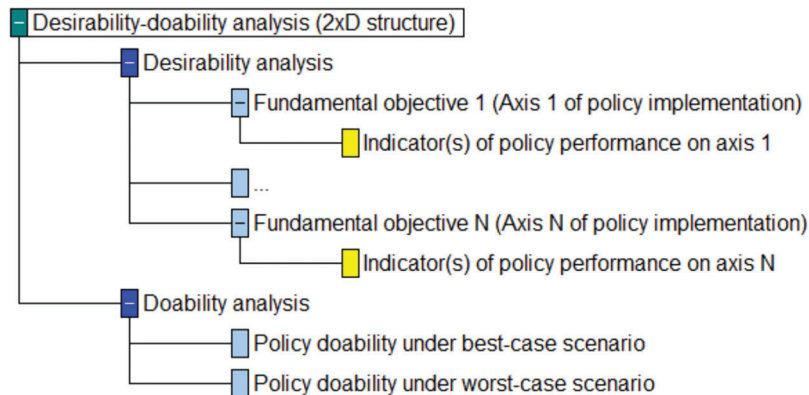


Fig. 1. Desirability–doability framework ($2 \times D$) model structure.

introduced in Section 1 through the development of a desirability–doability analysis as structured in Fig. 1:

Phase I structuring. The public policy context is characterized in terms of the actors involved and their fundamental objectives, the type of policies to improve the current situation on the objectives and how to appraise these improvements, and the barriers that may compromise their implementation and effectiveness. Contrasting scenarios of plausible futures are constructed to capture non-controllable external conditions that may affect, negatively or positively, both the effectiveness of policies and the frequency and magnitude of their barriers.

Phase II desirability. Deals with the measurement of overall desirability by constructing a group multicriteria model through a sequence of activities to measure objective-specific desirability and weight the fundamental objectives.

Phase III doability. Deals with the measurement of policy doability, under contrasting scenarios, to account for non-controllable external conditions affecting the seriousness of barriers to policy implementation.

Phase IV $2 \times D$ analysis. Facilitates the desirability–doability analysis of policies by using visual interactive graphs, in which each policy is represented by its measures of overall desirability and doability. The stability of the position of each policy in the graph is analyzed through sensitivity and robustness analyses.

Figure 2 outlines the $2 \times D$ framework. Its four phases run in face-to-face social processes, under the principles of process consultation (Schein, 1999) and in view of requisite decision modeling (Phillips, 1984): Phase I in facilitated workshops (Franco and Montibeller, 2010) with a broad panel of stakeholders, and Phases II to IV in computer-assisted decision conferences (Phillips, 2007; Parnell et al., 2013) with a core group of policymakers detached from the panel. Interaction protocols are followed by the facilitator in each phase, to help the group in “looking deeper into the subject, exploring, interpreting, debating and even arguing” (Roy, 2010, p. 77). Most modelling activities are supported in the M-MACBETH (Bana e Costa et al., 2017) software that implements the MACBETH multicriteria decision analysis approach (Bana e Costa et al., 2012). MACBETH (Measuring Attractiveness by a Categorical Based Evaluation Technique) requires only qualitative

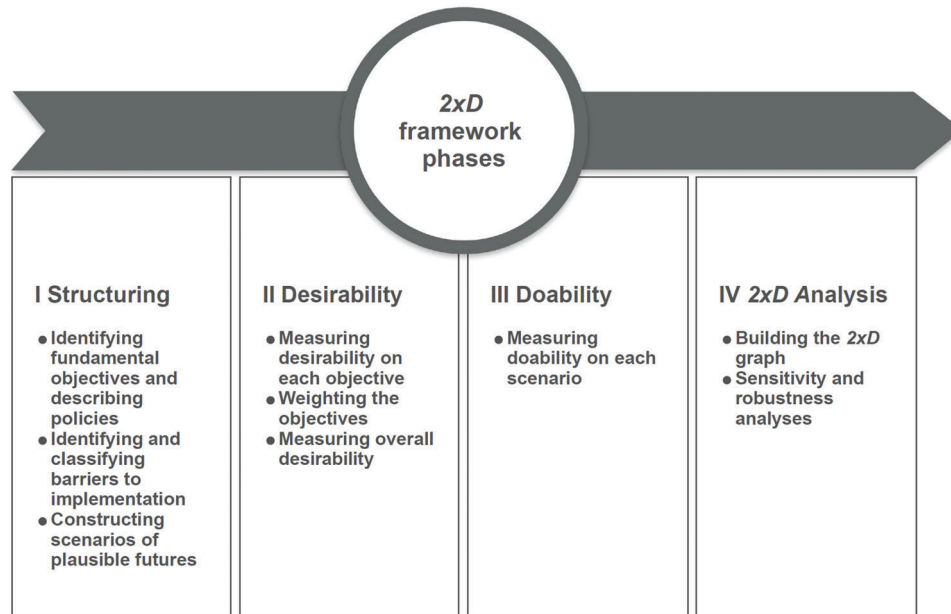


Fig. 2. Phases and activities of the $2 \times D$ framework.

(non-numerical) judgements of difference in attractiveness to quantify the relative attractiveness of policies, as overviewed in Section 2.6. “Attractiveness” is used in *lato sensu* in $2 \times D$, as it can refer to measuring the desirability (in Phase II) or the doability (in Phase III) of policies.

2.2. Phase I: Structuring

2.2.1. Activity I.1. Identifying fundamental objectives and describing policies

The “divide and conquer” strategy, mentioned in Section 1, requires a structuring protocol to identify the objectives that determine policy desirability. According to value-focused thinking (Keeney, 1992), the fundamental objectives that capture actors’ values (“what you hold to be of worth, useful, and desirable”; Hammond et al., 1998, p. 230) are first isolated and then used to drive the creation of new policies or make existing ones better means to achieve the designated fundamental objectives. In practice, each fundamental objective defines a policy intervention axis, and policies may vary from target-oriented, that is, designed for intervening on a single specific axis, to broad-spectrum policies that combine actions intended to affect several axes. In $2 \times D$, the facilitator starts by presenting and discussing an initial list of objectives with the panel, followed by groupwork to separate “means” from “ends”—by questioning the panel about why objectives in the list are important—until the fundamental objectives emerge. This activity can be facilitated by using collaborative groupwork methods (e.g., 6-3-5 brainwriting; Rhorbach, 1969) and problem structuring interaction protocols to construct group cognitive and oval maps (Ackermann et al., 2001; Eden, 2004), improving individual knowledge and developing a shared group understanding of the key objectives (as done, e.g., in Bana e Costa et al., 1999, 2006; Fernandes and Ferreira, 2020).

Once the set of fundamental objectives is agreed upon, policies should be characterized by the measures that compose them and by factual information about how well they are expected to contribute to changing the current situation (the *status quo*, *SQ*) on each fundamental objective. M-MACBETH allows a tree structure to be built (in the form illustrated in Fig. 1) with the set of fundamental objectives and their performance indicators resulting from Phase I. Policies and their performances can also be inputted.

This information will be later presented to the group in Phase II to stimulate the expression of desirability judgments, under conditions of full implementability. A descriptor of performance is defined by a (continuous or discrete) range of plausible performance levels in order to help analyze how well the policies perform on an objective. A peculiarity in $2 \times D$ is that one of the levels of the descriptor serves to characterize the *SQ* and another to characterize a “good” performance. These two levels are operational references used to develop better or new policies. In many cases, however, there is not a single measure of performance, and a multidimensional descriptor of a fundamental objective must be constructed by combining several performance indicators, which are usually means to achieve the objective. In contexts of non-spatial policies, that is, when their consequences are not spatially determined, the $2 \times D$ protocol for constructing a multidimensional descriptor for an objective adapts the procedure proposed by Bana e Costa and Beinart (2005): first, a small number of performance levels is selected in each indicator, which are combined to form performance profiles across the several indicators, and then unfeasible profiles are eliminated, and the remaining ones are rank-ordered to form a multidimensional performance scale for the objective. Finally, for each policy, it is then possible to associate the multidimensional profile that the panel selects as appropriate to represent how well the policy meets the objective.

With spatial policies, however, a particular protocol is necessary to address Challenge 6 (see Section 1) and overcome the recognized difficulty of anticipating geographically dispersed consequences (Simon et al., 2014; Keller and Simon, 2019). The spatial structuring protocol of $2 \times D$ focuses on the indicators to identify “critical situations” across the geographical units. Inspired by the process suggested by the World Health Organization and WHO Centre for Health Development (Kobe Japan) (2010), a critical situation is identified where the *SQ* has already achieved thresholds of urgency for intervention (e.g., maximum acceptable levels of emissions of pollutants set in regulations). A “good” policy for a fundamental objective can then be defined as a policy that would address and resolve all current critical situations across all geographical units on that objective. A policy desirability judgment will therefore be expressed for each fundamental objective, in terms of how well a policy improves the *SQ* over the territory, according to the number of current critical situations the panel expects the policy to resolve. This protocol also accommodates the negative side effects of object-oriented policies that may involuntarily contribute to generate additional critical situation(s) on other objectives.

2.2.2. Activity I.2. Identifying and classifying barriers

Barriers to implementation are also discussed with the experts in the workshops. Identifying barriers is crucial to policymaking in many sectors. For example, May et al. (2008) point out that a review by the European Conference of Ministers of Transport, in 2002, of the recommendations of a previous study on improving public transport in Europe, “concluded that, while the 1995 recommendations were broadly accepted, the implementation of such strategies was ‘more easily said

than done'. It highlighted as the principal barriers poor policy integration and coordination, counterproductive institutional roles, unresponsive regulatory frameworks, weaknesses in pricing, poor data quality and quantity, limited public support and lack of political resolve" (p. 328). As a rule of thumb, it is useful to distinguish fundamental from secondary barriers (Good et al., 2017). Depending on the context, the political, economic, social, technological, environmental, legal structure (Johnson et al., 2009) may be used to organize the collection of several types of barriers or impediments.

2.2.3. Activity I.3. Constructing scenarios of plausible futures

Phase I also deals with the construction of coherent scenarios of possible futures. $2 \times D$ follows this definition of scenarios: "scenarios primarily have a temporal property rooted in the future and reference external forces in that context; scenarios should also be possible and internally plausible while taking the proper form of a story or narrative description; scenarios seem to exist in sets, and the scenarios that inhabit those sets are systematically prepared to co-exist as meaningfully different alternatives to one another" (Spaniol and Rowland, 2019). $2 \times D$ proposes to develop two contrasting scenarios (worst-case and best-case) to bind the range of uncertainties that eventually affect the desirability and doability of policies. They are "narrative and qualitative descriptions and are best described using the concept of *mise-en-scène*" (Karvetski and Lambert, 2012). The three-stage scenario-building sociotechnical approach proposed in Alvarenga et al. (2019) is followed: Stage (i) identification of potential drivers relevant to the evolution of population health (PH) inequalities, Stage (ii) validation of drivers and generation of scenario structures, and Stage (iii) validation of scenario structures and generation of scenario narratives. This approach is rooted in the extreme-world method presented and applied by Goodwin and Wright (2014) (see also Shar et al., 2011).

2.3. Phase II: Desirability

2.3.1. Activity II.1. Measuring desirability on each objective

The facilitator of the $2 \times D$ decision conferences can use several tools to drive the elicitation of group qualitative judgments, in each objective-node of the tree. As explained in Section 2.6, the MACBETH technique derives quantitative scores for the policies from these judgments, which discussion allows the group to assign to each policy p a desirability score $v_j(p)$, on each fundamental objective j ($j = 1, \dots, n$).

The protocols designed for eliciting desirability judgments adapt the qualitative pairwise comparison elicitation mode of MACBETH, which introduces several qualitative categories of "difference of desirability," from no to extreme difference common to stimulate the comparison of policies for a fundamental objective. The questioning mode works as follows for a given pair of policies: assume there are no barriers affecting policy implementation, that is, all policies can be fully implemented; is there no difference in desirability, or is the difference in desirability between the two policies very weak, weak, moderate, strong, very strong, or extreme (or a sequence of these)? Each time a judgment is agreed upon, the consistency of all the judgments thereto agreed is verified, and suggestions are offered to resolve inconsistencies. When there is no inconsistency, a MACBETH quantitative scale of desirability scores on the objective is derived, respecting all qualitative judgments and their relationships. The second part of the protocol is devoted to validating the scale

Table 1
MACBETH-based three-step protocol for measuring the desirability of policies on each objective

Assume all policies can be fully implemented:

1. Use the MACBETH voting procedure to help the group qualitatively judge the desirability of each policy when compared to no change of the status quo (SQ), giving rise to a ranking
2. Use the MACBETH voting procedure to help the group qualitatively judge the difference in desirability between each two consecutive policies in the ranking
3. Along Steps 1 and 2, upload the group judgments being elicited to M-MACBETH, to test them for consistency, and at the end, derive the respective MACBETH desirability scale and present it to the group as a thermometer scale, already anchored in two fixed references: a “good” policy for the respective objective (with a desirability score of 100) and a policy making no contribution to improving the SQ (obviously, with a null desirability score); then, discuss the suggested scale with the group, and adjust it, if necessary, until final single-objective desirability scores for the policies are agreed upon

with the group, starting by taking the smallest numerical difference between scores as the unit of difference in desirability and then validating the ratios between each of the other differences and the unit.

In practice, asking the group to judge the differences of desirability for all pairs of policies is not strictly necessary in order to derive quantitative scores. Resorting to the MACBETH voting procedure can also be very helpful in significantly decreasing the elicitation time and getting the group aligned as in the cases reported in Bana e Costa et al. (2014) and Mateus et al. (2017). Adopting this protocol in $2 \times D$ involves eliciting only a few group desirability judgments for each objective. In short, the facilitator can follow the three-step sequence in Table 1.

The MACBETH voting procedure, used in both Steps 1 and 2, to facilitate the process of eliciting group judgments, develops in three steps: first, individual qualitative judgments are expressed by the members of the group, by “voting” in one MACBETH category, then the participants’ arguments justifying their individual votes are discussed, giving rise to group knowledge in the light of which a second round of individual voting is launched, if deemed convenient in order for the judgments to converge. Achieving consensus within the group is desirable, although a group compromise judgment derived from applying a (previously agreed) majority rule is acceptable. Firm minority disagreements are recorded during the application of MACBETH voting, and if they continue, they are later subject to sensitivity analysis.

2.3.2. Activity II.2. Weighting the objectives

The way in which the desirability scores of a policy contribute to its overall desirability is measured using relative weights assigned to the fundamental objectives. One can then measure the policy’s overall desirability by multiplying each objective-specific desirability score by the respective weight and summing these products across the objectives. In order to conduct the group weighting process, the facilitator must follow a protocol that avoids the critical mistake (the “most common” one; Keeney, 1992) of directly weighting the objectives in terms of relative importance. This is overcome in $2 \times D$ by focusing the group on comparing the “good” policies for each two objectives, in terms of differences in desirability, following the five-step qualitative weighting protocol of Table 2. It adapts the MACBETH protocol proposed in Bana e Costa et al. (2012) (Section 3.3) and, as in

Table 2
MACBETH-based five-step qualitative weighting protocol

1. Start by presenting the list of “good” policies for the several fundamental objectives to the group
2. Help the group judge which of the “good” policies is the most desirable, in terms of contributions to improving the SQ , and then qualitatively judge the difference in desirability between the most desirable good policy and each of the other “good” policies
3. Based on the judgments elicited in Step 2, help the group to rank-order the “good” policies from the most to the least desirable, in terms of contribution to improving the SQ , and then to qualitatively judge the difference in attractiveness between each two consecutive “good” policies in the ranking
4. Help the group to qualitatively judge how desirable the most desirable and the least desirable of the “good” policies are, in terms of contributions to improving the SQ
5. During Steps 1 and 4, upload the group judgments being elicited to M-MACBETH to test them for consistency, and at the end, derive the respective MACBETH weighting scale and present it to the group with the weights already summing up 1; then, discuss the suggested weights with the group and adjust them, if necessary, until final weights for the fundamental objectives are agreed upon

classic quantitative swing weighting protocol (von Winterfeldt and Edwards, 1986), asks the group to take into consideration both how big the performance gap is between “good” and the SQ , and “how much you care about it” (Phillips, 2014). Voting rounds can always be used to reconcile individual judgments at any elicitation step.

2.3.3. Activity II.3. Measuring overall desirability

M-MACBETH helps the facilitator to build a shared multicriteria additive desirability model with the group, in order to measure the overall relative desirability of policies - see Model (1). Simplicity is a core advantage of this model. Mathematically, the overall desirability score $v(p)$ of a policy p , which indirectly measures the overall desirability of p , in the eyes of the group, is given by:

$$v(p) = v(p) - v(SQ) = \sum_{j=1}^n k_j [v_j(p) - v_j(SQ_j)] = \sum_{j=1}^n k_j v_j(p) \quad (1)$$

where j designates a fundamental objective j ($j = 1, \dots, n$), k_j is the weight assigned to objective j (with $k_j > 0$ and $\sum_{j=1}^n k_j = 1$), and $v_j(p)$ is the desirability score of p on j - with $v_j(SQ_j) = 0$, where SQ_j represents the outcome of a policy that does not change the SQ on j . The objective-specific desirability score of 100 is arbitrarily assigned to the “good” policy ($good_j$) for each objective j , that is, $v_j(good_j) = 100$ ($j = 1, \dots, n$). Each product $k_j v_j(p)$ in Model (1) measures the contribution of the performance of policy p on objective j to the overall desirability of p .

The theoretical conditions of additivity (Dyer and Sarin, 1979, Smith and Dyer, 2021) implicit in Model (1) are constructively taken in $2 \times D$ as working hypotheses in building the model.

2.4. Phase III: Doability

A MACBETH-based protocol is also used in $2 \times D$ for eliciting doability judgments, in Phase III, while considering barriers to their implementation. Following the steps of the voting protocol in Table 3, the elicitation is made separately under each of the two contrasting scenarios previously

Table 3

Adapted MACBETH four-step protocol for evaluating policies' doability

-
1. Start by presenting the worst-case and best-case scenarios to the group and discuss them to ensure all participants apprehend them well, namely, how they differ in terms of describing possible futures
 2. Take one of the two scenarios and, with the group focused on the selected scenario, confront them with the questions: in the face of the envisaged seriousness of the barriers to implementation identified for each policy, how do you judge the doability of the policy on this scenario: null, very weak, weak, moderate, strong, very strong, or extreme (or a sequence of categories)? For two given policies, how do you judge their difference of doability in this scenario?
 3. Use the MACBETH voting procedure: for each policy, start by eliciting individual answers to the first question enounced in (2), launch a discussion around them, inviting participants to justify their judgments to the group; eventually, launch a second round allowing the revision of initial individual judgments, in light of the knowledge acquired from the discussion; following a majority rule, suggest a group judgment and discuss and revise it with the group until a compromise doability judgment is agreed. Ask the group to rank-order the policies by decreasing relative doability, and answer the second question enounced in (2), for pairs of consecutive policies in the ranking and following an elicitation sequence as in (2)
 4. Apply MACBETH to the group judgments agreed in (3), discuss the derived doability scale with the group, and if necessary, adjust it
 5. Repeat Steps 2 to 4 for the other scenario
-

constructed (worst-case and best-case), starting by updating the barriers. Technically, the scale of doability varies from 0 (null doability) to 100 (extreme doability). A null doability is assigned to a policy affected by barriers that would make the group consider it unrealistic to implement, whereas extreme doability corresponds to a case with no, or only inconsequential, barriers.

2.5. Phase IV: $2 \times D$ analysis

2.5.1. Activity IV.1. Building $2 \times D$ graphs

The achievement of group agreement (consensually or by majority) on the overall desirability and doability of policies marks a turning point in the $2 \times D$ socio-technical process, allowing decision-conferencing to move forward to the cross-analysis of desirability and doability results. This is done, in each of the two previously constructed scenarios, with the visual support of a $2 \times D$ visual interactive graph, in which each policy is represented and analyzed by its measures of overall desirability and doability, produced with the XY Map tool of M-MACBETH. A classification of policies in four categories is useful for policymaking: “pearls” (policies with high desirability and easy to implement), “oysters” (those with high desirability but difficult to implement), “bread and butter” (easy to implement but of low desirability), and “white elephants” (low desirability and difficult to implement). These policy categories were first used in Bana e Costa et al. (2014). For example, knowing the barriers faced by oysters facilitates the design of measures to mitigate or break the barriers, in view of converting oysters into pearls.

2.5.2. Activity IV.2. Sensitivity and robustness analyses

M-MACBETH incorporates visual interactive tools to develop extensive sensitivity and robustness analyses, motivated by “what-if” questions, in order to address the different types of uncertainty

phenomena that can affect the results of the multicriteria desirability model. Sensitivity analysis is classic in decision analysis textbooks (see, e.g., Clemen, 1996, Chapter 5) and is limited to the effects on the overall desirability scores caused by varying only one type of parameter in Model (1) at a time, either a single-objective score or a weight. Dealing with the two types simultaneously requires a robustness analysis, which can be operationally defined as an extension of classical sensitivity analysis to allow for simultaneous variations of several model parameters (see, e.g., Bertsch et al., 2007). There are various types of robustness analyses, and “to carry out meaningful robustness analysis, it must be made clear which unknown quantities and parameters are to be considered for the analysis, and what the variation or uncertainty is reflecting” (Aven, 2013, p. 2088). In M-MACBETH, robustness analysis is developed on the concept of “additive dominance” (Bana e Costa et al., 2012). A policy additively dominates another policy if it is always found to be more desirable than the other—that is, the difference between the multicriteria desirability scores of the former and the latter given by Model (1) is always positive—under some constraints to the variation of input data under different scenarios. For example, under the worst-case scenario, it may be that some situations become critical or others already critical become even more critical, in a part or all over the territory; consequently, a policy may decrease its performance in terms of overcoming critical situations. The converse is true under the best-case scenario. Operationally, these phenomena can be incorporated in MACBETH modeling throughout intervals of variation defined for the input data.

Formally, robustness analysis can be presented as follows. Let p and q be any two policies under consideration, $\Delta(p, q) = v(p) - v(q) = \sum_{j=1}^n k_j [v_j(p) - v_j(q)]$ a measure of the difference of overall desirability between p and q given by Model (1). Let U be a set of constraints modeling some uncertain information. Also, let $\text{Min}\Delta(p, q)$ and $\text{Max}\Delta(p, q)$ be, respectively, the minimum and maximum values of $\Delta(p, q)$ under U . A mathematical programming algorithm (De Corte, 2002) is implemented in M-MACBETH to calculate $\text{Min}\Delta(p, q)$ and $\text{Max}\Delta(p, q)$ under several types of U —such as, for example, intervals of variation of desirability scores v_j on some objectives and/or of their weights k_j —allowing a robustness analysis, for each pair of policies, as follows:

$$\begin{cases} \text{Min}\Delta(p, q) \geq 0 \text{ and } \text{Max}\Delta(p, q) > 0, & p \text{ additively dominates } q \\ \text{Min}\Delta(p, q) < 0 \text{ and } \text{Max}\Delta(p, q) \leq 0, & q \text{ additively dominates } p \\ \text{Min}\Delta(p, q) < 0 \text{ and } \text{Max}\Delta(p, q) > 0, & p \text{ and } q \text{ are incomparable} \end{cases} \quad \text{under } U \quad (2)$$

Incomparability means that the results of the additive model built are not stable for p and q under the conditions of uncertainty defined by the constraints.

2.6. The MACBETH technique

Used extensively to support the activities in Phases II and III of $2 \times D$, MACBETH was first proposed in this journal as a novel “interactive path toward the construction of cardinal value functions” (Bana e Costa and Vansnick, 1994). It is an alternative non-numerical value elicitation technique to numerical techniques such as direct rating (von Winterfeldt and Edwards, 1986). Although founded on the principles of value-difference measurement (Dyer and Sarin, 1979), these two techniques “are not psychologically equivalent” (Fasolo and Bana e Costa, 2014). MACBETH

“is perceived as a convenient way to express value judgements by lowering cognitive load” (Angelis and Kanavos, 2017, p. 150) because it only requires the expression of qualitative judgments. The last update of MACBETH (Bana e Costa et al., 2012) kept the original qualitative elicitation protocol based on the seven qualitative categories of difference in attractiveness, from no to extreme difference (as described under Protocols II in Section 2.2), although the protocol has since been extended to accommodate judgments of hesitation or disagreement between consecutive categories (e.g., the difference of attractiveness is “moderate or strong”). The fundamental idea used to derive a numerical value scale from a set of consistent judgments is straightforward and twofold:

1. If the difference of attractiveness was judged...

extreme: assign to it a numerical value of 6, or, if not possible, greater than 6

very strong: assign to it a numerical value of 5, or, if not possible, greater than 5

strong: assign to it a numerical value of 4, or, if not possible, greater than 4

moderate: assign to it a numerical value of 3, or, if not possible, greater than 3

weak: assign to it a numerical value of 2, or, if not possible, greater than 2

very weak: assign to it a numerical value of 1, or, if not possible, greater than 1

null (indifference): assign to it a numerical value of 0,

such that

2. if one difference of attractiveness was judged more intense than another, then the numerical value assigned to the former judgment must be greater than the numerical value assigned to the latter (a condition of order preservation).

These conditions can be mathematically formulated in a linear programming problem, which solution involves associating a numerical score with each qualitative judgment elicited. Conceptually, let X be the set of policies, $v(x)$ the score assigned to policy x of X , and x^+ and x^- two policies of X such that x^+ is at least as attractive as any other element of X and x^- is at most as attractive as any other element of X . The following formulation, labeled LP-MACBETH in Bana e Costa et al. (2012), already allows for dealing with hesitation judgments expressed by two or more consecutive categories of difference of attractiveness, C_1 (very weak) to C_6 (extreme).

$$\text{Min}[v(x^+) - v(x^-)] \quad (3)$$

Subject to

1. $v(x^-) = 0$ (arbitrary assignment)
2. $v(x) - v(y) = 0, \forall (x, y) \in C_0$ (indifference)
3. $v(x) - v(y) \geq i, \forall (x, y) \in C_i \cup \dots \cup C_s$ with $i, s \in \{1, 2, 3, 4, 5, 6\}$ and $i \leq s$
4. $v(x) - v(y) \geq v(w) - v(z) + i - s', \forall (x, y) \in C_i \cup \dots \cup C_s$ and $\forall w, z \in C_{i'} \cup \dots \cup C_{s'}$, with $i, s, i', s' \in \{1, 2, 3, 4, 5, 6\}, i \leq s, i' \leq s'$ and $i > s'$.

Conditions 2 to 4 are conditions of order preservation (COP) that ensure the ranking of the elements (COP 2 and 3) and the order between differences of attractiveness (COP 4). Note that no condition is imposed between judgments of the same category, to which can be assigned the same

or different numerical scores. That is, the qualitative categories can be represented as a sequence of non-overlapping intervals of real numbers, with the objective function contributing to minimize the size of each category, and, if possible, reduce them all to single numbers. It is important to emphasize that it is not necessary to make all the $m(m - 1)/2$ qualitative pairwise comparisons possible within a set of m policies. As said, this allows a significant reduction in the time spent on the elicitation process, as many missing judgments can be derived by transitivity. Nevertheless, and although the minimum number of judgments necessary to find a scale with LP-MACBETH is $(m - 1)$, it is recommended that more judgments are elicited to allow for consistency checks. For example, the number of judgments elicited through the MACBETH voting procedure is at least $(2m - 3)$.

If there is no solution to LP-MACBETH, the set of elicited judgments is not consistent and should be revised. The suggested changes to elicited judgments, mentioned above, result from other technical programs presented in Bana e Costa et al. (2005), which shows how the minimum number of changes necessary to overcome inconsistency can be found. When consistency is verified, a unique solution (called the MACBETH basic scale) is always proposed, if necessary, using supplementary programs when there are multiple optimal solutions for LP-MACBETH (see Bana e Costa et al., 2005). The complete set of programs is implemented in the M-MACBETH decision support system; however, it is worthwhile emphasizing that the basic solution given by resolving LP-MACBETH can be determined “by hand,” following the procedure presented in Bana e Costa et al. (2012), in which application for a small consistent set of judgments can be easily shown to the group, therefore increasing modeling transparency. The resulting basic MACBETH scale should then be discussed with the group, by comparing differences in scores (intervals in a thermometer scale displayed in M-MACBETH—see example in Fig. 5) and adjusting one or more scores, if necessary, until an agreement is reached on a final interval scale (i.e., a numerical scale unique up to a positive linear transformation). M-MACBETH also visually displays the interval within which each score can vary without violating any COP, to facilitate the adjustment of scores. This essential—yet often ignored—(cardinal consistency) checking is the step in value-difference measurement where MACBETH and direct rating procedures become “technically equivalent” (Fasolo and Bana e Costa, 2014). At the end of the day, MACBETH offers a way to avoid starting a judgmental elicitation process by asking directly for numerical ratings that a group can find “hard to answer” (von Winterfeldt and Edwards, 1986).

3. Testing the $2 \times D$ framework with the support of a real-world case

3.1. The Lisbon case

The $2 \times D$ framework was tested in the evaluation of urban health policies, with performances spread over the Lisbon municipality in 24 policy units (listed in the table at the right in Fig. 4). Several city policymakers were actively engaged in developing the activities of the four phases proposed in Section 2. A panel of 32 regional and local stakeholders was constituted to participate in structuring workshops. Care was taken when selecting policy agents from different sectors (local and regional government, charities and other non-profit and non-governmental organizations, public health, and healthcare services) and with different viewpoints concerning the benefits of municipal policies to health. An evaluation group of 16 policymakers was selected from this panel,



Fig. 3. Lisbon decision conferencing in progress.

with the help of the city councillor responsible for health and social affairs, to participate in the decision conferencing process to develop the activities of Phases II to IV. As all the group members participated in the structuring workshops, the diversity of perspectives that emerged during Phase I was present in the decision conference.

In total, the Lisbon $2 \times D$ process comprised three working days of face-to-face interaction. Two half-day panel workshops took place on the afternoons of 26 November 2016 and 20 February 2017, devoted to the structuring activities of Phase 1 of the $2 \times D$ framework (see Section 3.2, viz., the identification of indicators and critical situations, upon which the policies were characterized (as detailed in Freitas et al., 2020). Two one-day group decision conferences then took place on 26 and 29 May 2017, each day involving four working sessions of two hours, in which the group developed the several activities of $2 \times D$ Phases II to IV. Support material with the conclusions of the workshops was sent to each member of the group on 19 May as an annex to the calling note for the decision conference. The Lisbon $2 \times D$ desirability–doability modeling process was developed on the spot with the group, supported by the projection of M-MACBETH on a big screen, in a conferencing room with the layout shown in Fig. 3. As remarked by Phillips

and Bana e Costa (2007): “Because the model is projected for all participants to see it as it is created, it is less likely to be perceived by participants as a ‘black box’, which helps to gain confidence in model results” (p. 54). Measuring the desirability of policies for each of the eight fundamental objectives occupied the three first working sessions of the first day (see Section 3.3.1), and the last session of the day was devoted to weighting the objectives (see Section 3.3.2). The first hour of the second conferencing day was devoted to an overview of the additive desirability model and the discussion of the overall desirability scores of the policies resulting from applying the model with the objective-specific desirability scores for the policies and the weights of the objectives set in the first day (see Section 3.3.3). The group then turned to measuring the doability of the policies under each of the two scenarios, which was finished at the end of the second working session (see Section 3.4). After lunch, the third working session was devoted to analyzing the desirability–doability graphs and discussing the classification of the 18 policies (see Section 3.5.1). The decision conference finished with the session on sensitivity and robustness analyses (see Section 3.5.2).

As said in Section 1, the case was intended to be a real-world proof of concept of the feasibility of implementing the $2 \times D$ framework proposed in Section 2. It was therefore important to get *ex-post* feedback from participants. A few weeks after the end of the second decision conference, three policymakers, who had participated in all the panel and group sessions, were invited to provide their thoughts on the process in individual interviews (see Section 4.2). They were the Lisbon city councilor, responsible for health and social affairs (the top municipal politician in terms of decision-making power), the municipal spatial planning chief (the top technical staff member of the municipal policymaking team), and an official from the regional health authority (a top administrator in the Lisbon Region).

3.2. Phase I: Structuring facilitated workshops

In the facilitated structuring workshops, the panel selected 28 relevant indicators across eight “health determinants” (discussed in Bana e Costa et al., 2022), which are the fundamental objectives included in the value tree at the left in Fig. 4. The value tree also highlights the three indicators of *physical environment* [PE]. The *SQ* in each objective was then characterized by the critical situations for the respective indicators, identified in each Lisbon policy unit. For example, the table at the right in Fig. 4 shows the critical situations on [PE] identified by the panel. Analyses and discussions around the *SQ*, across the eight fundamental objectives, informed the review of 18 policies (listed in Table 4).

Policies cover a variety of policy domains, from promoting lifelong education to the reduction of air pollution and noise, each one integrating several municipal measures. A few are horizontal or cross-sectorial policies, being beneficial to more than one objective, as shown in Table 4. *Cohesion* is a horizontal policy, whereas most other policies are vertical, that is, target-oriented to a particular objective. Except for *road safety*, all objectives are tackled by more than one policy, with *built environment* being the objective impacted by the highest number of policies (seven). Note that, as shown in the last line of Table 4, at least one of the 18 policies has a “good” performance on each objective.

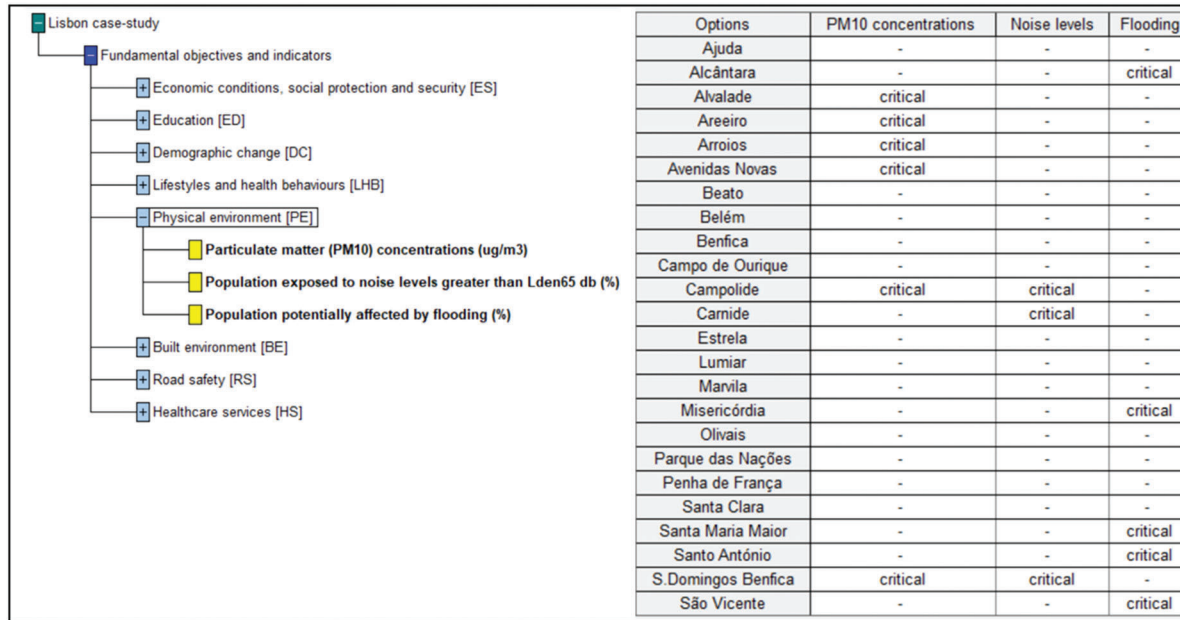


Fig. 4. Value tree of the Lisbon case with fundamental objectives, indicators of the physical environment (PE) objective, and table of policy units with critical situations that are options for policy intervention (table adapted from Freitas et al., 2020).

Policies were also analyzed in terms of expected barriers to their implementation and identified for the measures of each policy. For example, with respect to [Polut], the creation of reduced emissions zones in the city, viewed as an important measure to reduce pollution, would have to face impediments related to social and individual resistance to change, in part linked to a shortage of individual financial resources to purchase more fuel-efficient vehicles, but also to inertia in adopting other forms of mobility; whereas the shortage of municipal investments funds could significantly affect other [Polut] measures, such as the renewal of the bus and tram fleet, the creation of 4455 new parking spaces near bus stops outside the city center, the reduction of public transport prices, and the purchase of 33 new metro trains. Often, the measures of a public policy should be inserted in legal policy instruments, the approval of which depends on a majority of favorable votes in the Lisbon municipal council. For example, creating reduced emissions zones depends on the approval of the Lisbon Air Quality Action Plan.

In Activity I.2, the scenarios used in the Lisbon case summarize two contrasting sets of unfolding events that could plausibly affect the evolution of health inequalities across European Union regions until 2030. Details of the construction of these European macro scenarios, and their thorough descriptions, can be found in Alvarenga et al. (2019). The use of scenarios helped the Lisbon group to judge the extent to which possible future events in Europe could affect the evaluation of Lisbon policies and the seriousness of barriers. The worst-case scenario, named *Failing Europe*, imagines that Europe plunges into a new, deeper, and long-lasting economic crisis, with health inequalities increasing across Europe. The best-case scenario, *Sustainable Prosperity*, assumes that health inequalities decrease across Europe.

Table 4
Fundamental objectives impacted by the policies

Policies	Fundamental objectives										Fundamental objective(s) which are fully addressed
	Economic conditions, social protection and security [ES]	Education [ED]	Demographic changes [DC]	Lifestyles and health behaviors [LHB]	Physical environment [PE]	Built environment [BE]	Road safety [RS]	Healthcare services [HS]	Objectives per policy		
[Cohesion]	•	•	•	•		•		•		5	[ES], [BE]
[QoL]			•					•		2	[DC], [HS]
[PHC]			•	•				•		3	–
[Educa]	•		•							2	–
[UseTrans]					•					2	–
[UrbReab]					•					2	–
[SchLeav]				•						2	[ED]
[SexLit]				•						2	[LHB]
[Housing]									•	1	–
[SoftMob]					•				•	2	–

Continued

Table 4
(Continued)

Policies	Fundamental objectives										Fundamental objective(s) which are fully addressed
	Economic conditions, social protection and security [ES]	Education [ED]	Demographic change [DC]	Lifestyles and health behaviors [LHB]	Physical environment [PE]	Built environment [BE]	Road safety [RS]	Healthcare services [HS]	Objectives per policy		
[Employ]	•								1		–
[Access]			•				•		2		–
[Polut]					•				1		[PE]
[Flood]					•				1		[PE]
[RoadSaf]									1	•	[RS]
[UrbReg]							•		1		–
[SocInt]									1		–
[EfeTrans]								•	1		–
Policies per objective	4	3	5	3	6	7	1	3			
“Good” policy for the objective	[Cohesion]	[SchLeav]	[QoL]	[SexLit]	[Polut], [Flood]	[Cohesion]	[RoadSaf]	[QoL]			

Note: A • (respectively, a •) in a cell means that the critical situations for the objective in the column are fully (respectively, partially) addressed by the policy in the line; an empty cell means no impact.

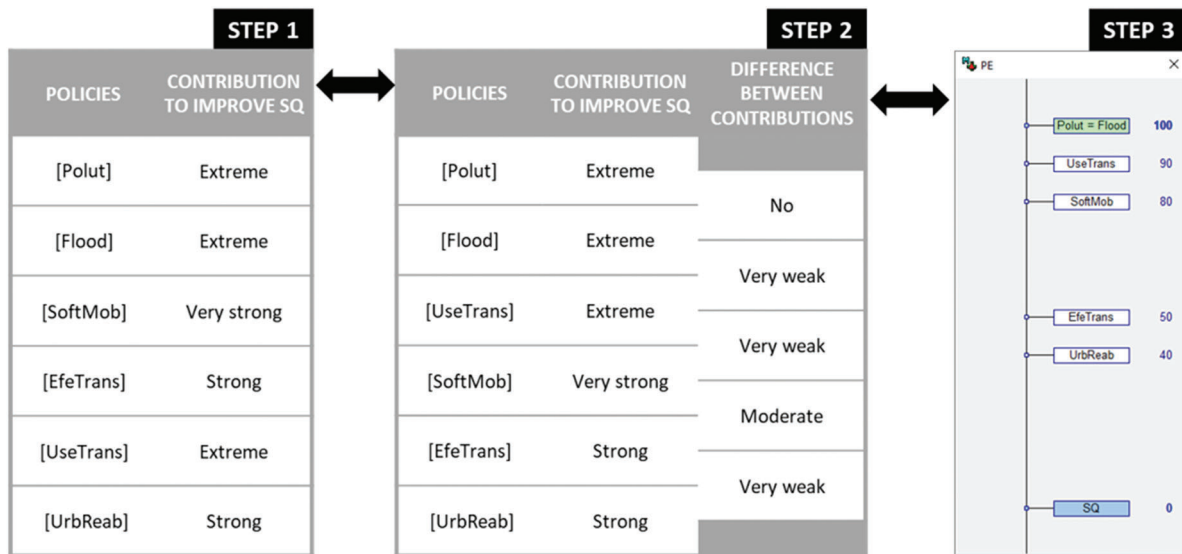


Fig. 5. Steps of the MACBETH protocol for evaluating the contribution of six policies to improving the SQ for physical environment [PE] (policy descriptions in Table 4).

3.3. Phase II: Evaluation decision conference

3.3.1. Measuring objective-specific desirability

The Lisbon decision conference started with the group application of the three-step protocol (Table 1) for evaluating the policies in terms of their desirability, for each of the eight fundamental objectives. Figure 5 depicts the process for the [PE] objective. From the set of six policies that contribute to this objective (see Table 4), the group judgments ranged between *extreme* (for both [Polut] and [Flood] policies) and *strong* (policy [UrbReab]) contributions to improving the SQ on [PE] (Step 1 in Table 1). After rank-ordering the policies accordingly to these judgments, the group pairwise compared every two consecutive policies in the ranking (Step 2 in Table 1). For example, they judged the difference of desirability between the policies [UseTrans] and [SoftMob] as *very weak* in terms of improving [PE]. In Step 3, the group analyzed and adjusted the proposed MACBETH scale, ensuring that it reflected the perceived value differences between policies. The same sequence was followed to set numerical value scores for all policies on the seven remaining objectives. During the use of the protocol, no situation emerged, for any objective, that could imply the non-verification of independence for the set of fundamental objectives.

3.3.2. Weighting the fundamental objectives

The second part of the Lisbon decision conference was devoted to weighting the eight fundamental objectives. Figure 6 details the application of the five-step qualitative weighting protocol (see Table 2). As can be seen in the table corresponding to Step 2 in Fig. 6, the [Cohesion] policy was

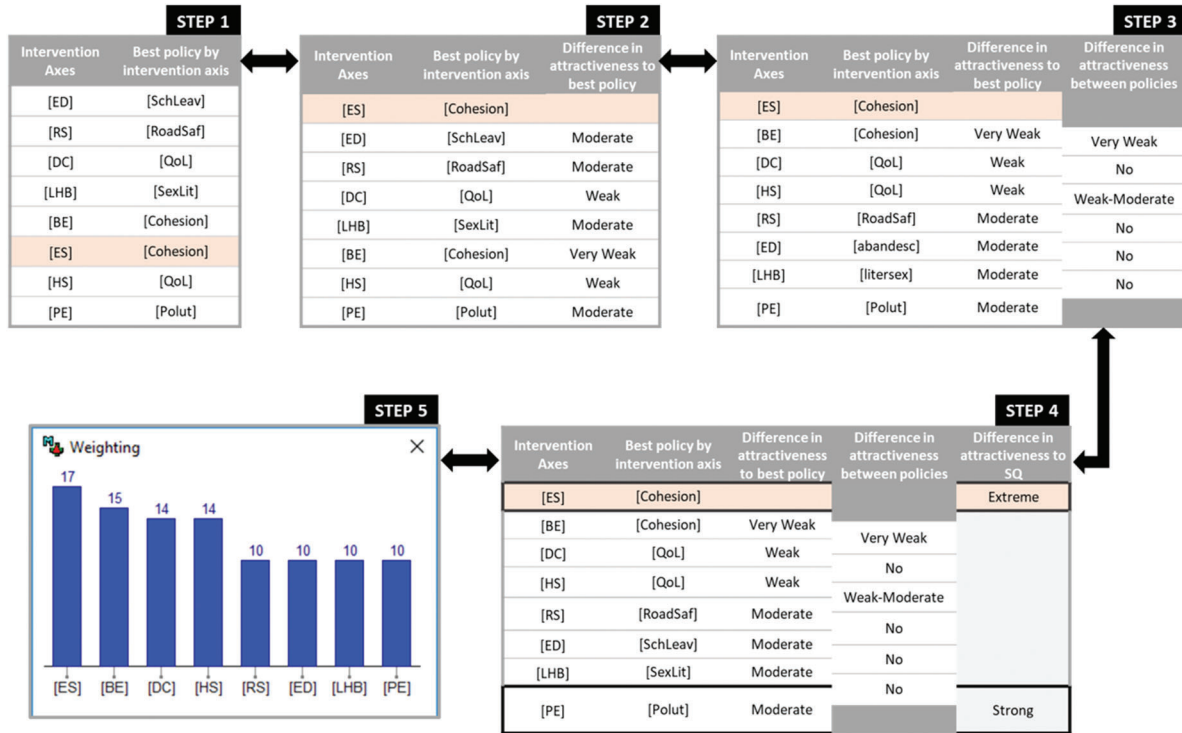


Fig. 6. MACBETH-based qualitative weighting procedure.

judged the most attractive among all the “good” policies, and the judgments between [Cohesion] and each one of the other “good” policies are shown in the same table. Next, the table for Step 3 displays the ranking of the “good” policies, and the judgments elicited for each two consecutive “good” policies in the ranking were added. Finally, the overall judgments elicited for the most and least attractive policies were added in the table for Step 4. The results of the weighting process, that is, the weights the group agreed to assign to the objectives, are displayed in percentages in the histogram for Step 5 in Fig. 6, which resulted from the elicited weighting judgments inserted in the matrix of Fig. 7. The judgments elicited in Step 2 were inserted in the first row of the matrix, those elicited in Step 3 form the first diagonal above the main diagonal of the matrix, and the last column was filled in with the judgments elicited in Step 4.

Although a compromise was agreed within the group on these weights, they were not consensual because one stakeholder from the healthcare sector was dissatisfied with the short difference between the relative weights of [HS] and [ED] and argued in favor of giving more importance to healthcare. Usually, “dissatisfaction on the part of the participants with elements of the model, or its results, drives the dialectic in the group, resulting in further changes to the model” (Phillips and Bana e Costa, 2007, p. 382). Nonetheless, a significant majority was formed to move the process forward with the weights of Fig. 6, with the facilitator’s promising to return, later on, to the argument of the divergent stakeholder, when analyzing the extent to which a higher weight given to the [HS] objective would affect the model results (see Section 3.2.3). Hesitations or differences

	[ES]	[BE]	[DC]	[HS]	[RS]	[ED]	[LHB]	[PE]	SQ
[ES]	no	very weak	weak	weak	moderate	moderate	moderate	moderate	extreme
[BE]		no	very weak	very weak	positive	positive	positive	positive	positive
[DC]			no	no	weak-mod	weak-mod	weak-mod	weak-mod	positive
[HS]			no	no	weak-mod	weak-mod	weak-mod	weak-mod	positive
[RS]					no	no	no	no	strong
[ED]					no	no	no	no	strong
[LHB]					no	no	no	no	strong
[PE]					no	no	no	no	strong
SQ									no

Consistent judgements

Fig. 7. MACBETH matrix of group weighting judgments (policies' description in Table 4).

in opinion among group members were accommodated by synthesizing group judgments in more than one MACBETH qualitative category—this is the case for the weak or moderate differences in Figs. 6 and 7.

3.3.3. Measuring overall desirability

A shared additive desirability Model (1) was built with the weights and objective-specific desirability scores obtained in Phase II. Its application to the 18 policies gave rise to the results displayed in Fig. 8, where the policies are listed by decreasing (overall) desirability. The group was not surprised by the significant gap in desirability (i.e., in added value to the *SQ*) between the horizontal policy [Cohesion] and the other policies, given that most are essentially target-oriented on one intervention axis: they are essentially vertical policies. One important conclusion was the need to add new measures to the vertical policies to make them more horizontal.

3.4. Phase III: Group appraisal of doability of policies for each scenario

In the part of the decision conference devoted to doability, the MACBETH-based four-step voting protocol (see Table 3) was used to appraise the doability of each policy, separately in each scenario. The group focused first on the worst-case scenario (*Failing Europe*), and after reaching an agreement, the facilitator moved the elicitation process forward to the best-case scenario (*Sustainable prosperity*). The number of individual doability judgments of the same category given by the participants was registered for each policy (as depicted in Table 5), enabling an immediate appraisal of the disparity of perceptions within the group. Participants were then invited to share the reasons behind their judgments and debated them in an interactive process in which they could change their initial individual judgments in light of new knowledge acquired during the discussion. It is worth highlighting that the conversation revealed on several occasions that the initial choices of some stakeholders were based on a mixed perception of doability and desirability, not on the former alone (as it methodologically should be). Every time this type of judgmental bias arose, the facilitator intervened to refocus the group.

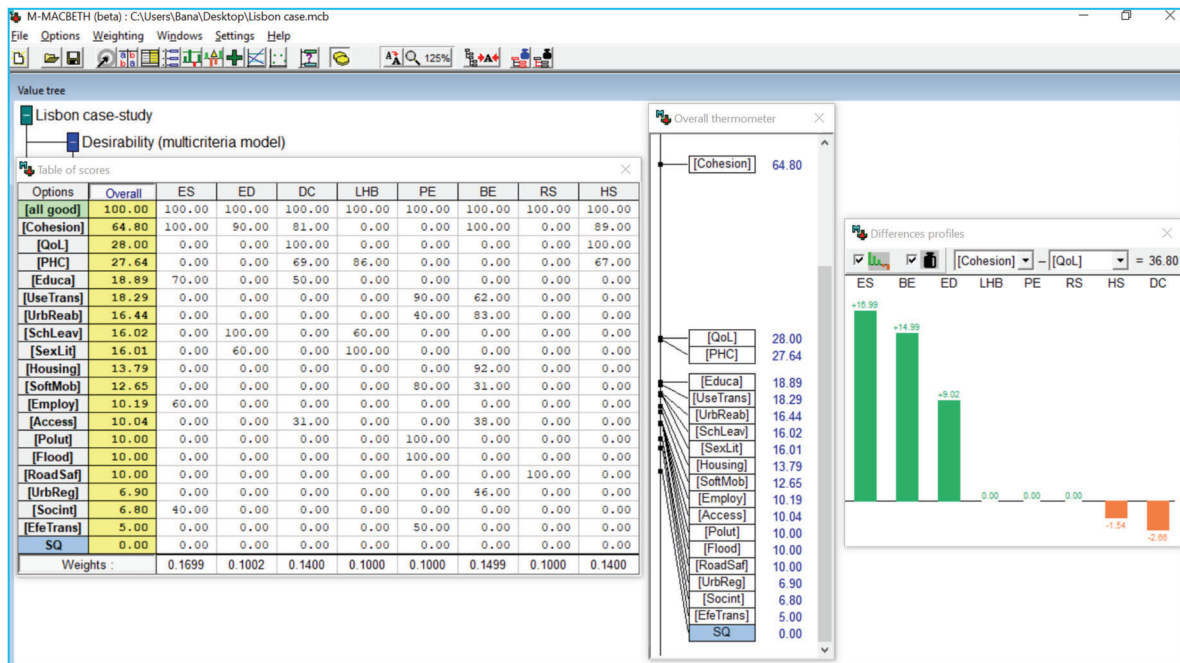


Fig. 8. M-MACBETH table of scores (with the objective-specific value scores and the multicriteria desirability scores of the policies, and weights of the objectives), the thermometer of (overall) desirability scores, and graph with the profile of weighted differences in desirability between the two first policies in the overall ranking, [Cohesion] and [QoL] (a green bar means that the respective objective is favorable to the former, whereas in a red bar, it is favorable to the latter) (policy descriptions in Table 4).

Table 5 also shows the compromise doability judgments subsequently agreed by the group, under each of the two scenarios. Differences in doability between judgments of consecutive categories were all taken as very weak, and judgments of the same category were all taken as indifferent. The policy doability scores proposed by M-MACBETH were discussed and adjusted until the group reached an agreement (see the final scores in Table 5). Longer discussion time was necessary, in general, for the worst-case scenario, compared to the best-case scenario, before the group were aligned. There was a larger dispersion of individual doability judgments for certain policies, raising political versus technical controversy issues. This was the case for [Polut], [SoftMob], [UseTrans], and [EfeTrans]. On the one hand, political arguments were often signaled in favor of higher doability, even under *Failing Europe*, for those policies already in implementation, or planned and budgeted for the medium-long term, and it was therefore easier for politicians to show continuity in the work being conducted. On the other hand, the lower doability of the same policies was justified by technical issues linked to their nature and by the expected escalation of their implementation barriers, in the face of which political wishes would not be enough to keep those policies in case of a longer and deeper economic crisis. In the end, it was clear that the group agreed that policies would be easier to implement under *Sustainable Prosperity* (strong to very strong group doability) than under *Failing Europe* (weak to strong doability).

Table 5
MACBETH process for measuring doability of policies for each scenario (policy descriptions in Table 4)

Policies	Doability in “FAILING EUROPE” scenario					Doability in “SUSTAINABLE PROSPERITY” scenario											
	Individual judgments					Individual judgments											
	No	Very weak	Weak	Moderate	Strong	Very strong	Extreme	Group judgment	Doability score	No	Very weak	Weak	Moderate	Strong	Very strong	Extreme	Group judgment
[Cohesion]				8	2	2		Moderate-Strong	58				2	9	1	Very Strong	83
[QoL]		1	5	3	3		Moderate	50				3	9		Very Strong	83	
[PHC]			3	5	4		Moderate	50				5	7		Strong-Very Strong	75	
[Educa]		1	4	4	3		Moderate	50				1	10	1	Strong	67	
[Use Trans]			2	2	4	4	Strong	67				1	10	1	Very Strong	83	
[UrbReab]			8	4			Weak	33				4	8		Very Strong	83	
[SchLeav]			1	7	1	3	Moderate-Strong	58				5	7		Strong-Very Strong	75	
[SexLit]			2	6	4		Moderate	50				4	8		Very Strong	83	
[Housing]			4	5	3		Moderate	50				6	6		Strong-Very Strong	75	
[SoftMob]			1	1	6	3	Strong	67		1		5	7		Strong-Very Strong	75	
[Employ]		5	1	5		1	Weak	33				3	9		Strong	67	
[Polut]		1	2	6	3		Moderate	50				11	1		Very Strong	83	
[Flood]				4	4	4	Strong	67				2	10		Very Strong	83	
[Access]			1	4	5	2	Strong	67				4	7	1	Very Strong	83	
[RoadSaf]				7	4	1	Moderate-Strong	58				3	9		Very Strong	83	
[UrbReg]			3	8	1		Moderate	50				7	5		Strong-Very Strong	75	
[SocInt]		2	4	2	3	1	Moderate	50				2	10		Strong	67	
[EfeTrans]			3	5	4		Moderate	50				1	11		Very Strong	83	

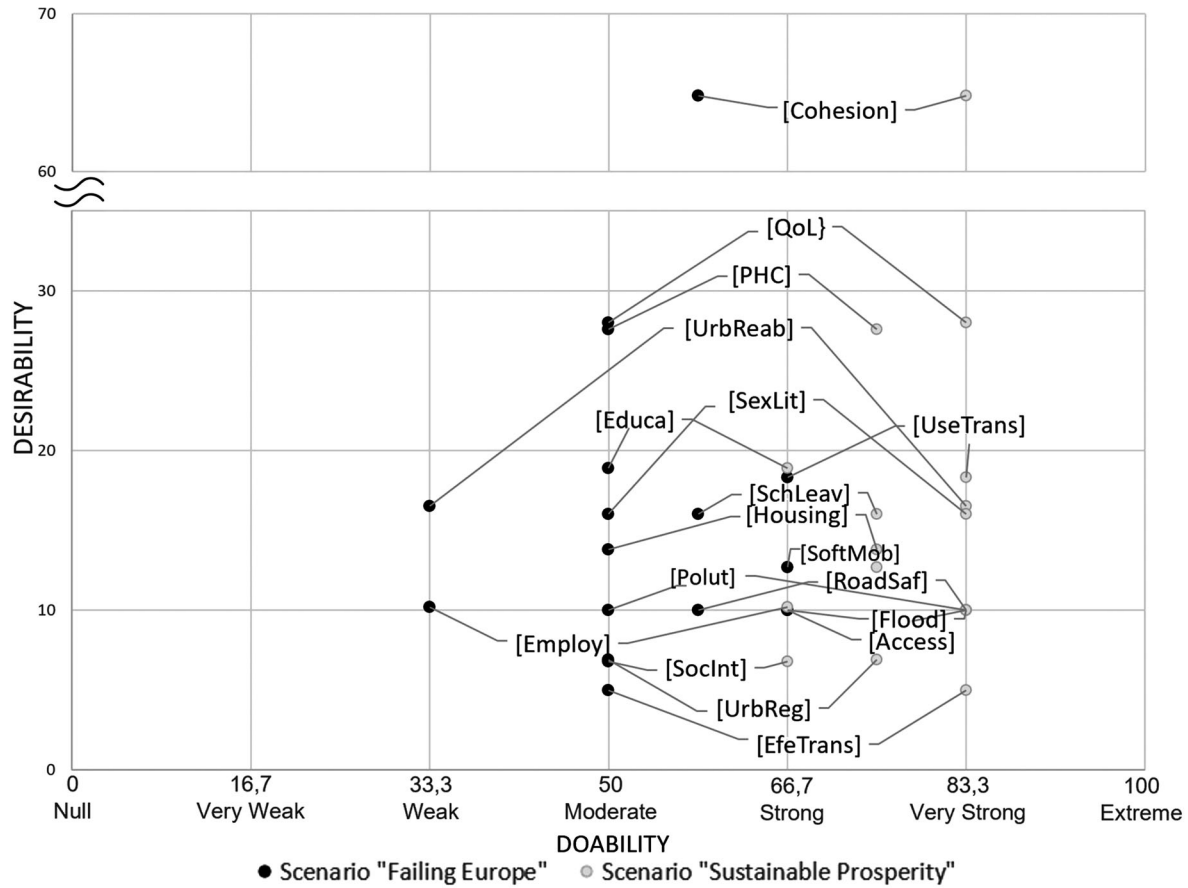


Fig. 9. Desirability–doability graph depicting the location of each policy under the two scenarios (policy descriptions in Table 4).

3.5. $2 \times D$ analysis

3.5.1. Building the $2 \times D$ desirability–doability graphs

Two separate $2 \times D$ visual interactive graphs were developed for the Lisbon case with the support of M-MACBETH. They were put together in the merging graph of Fig. 9, which was presented to the group for discussion. While the implementation of some policies (e.g., [Employ]) was found to be significantly affected by the external context, others were found to have high doability independent of the context (e.g., [UseTrans]). The group agreed that all policies except [Cohesion] needed to be enriched with new actions/measures capable of improving both their desirability and doability, thereby anticipating significant effects in municipal resource allocation. Indeed, one measure (at least) was identified in each policy as potentially requiring alternative sources of funding in the *Failing Europe* scenario. It was not possible to design new measures during the decision conference, although some participants made suggestions.

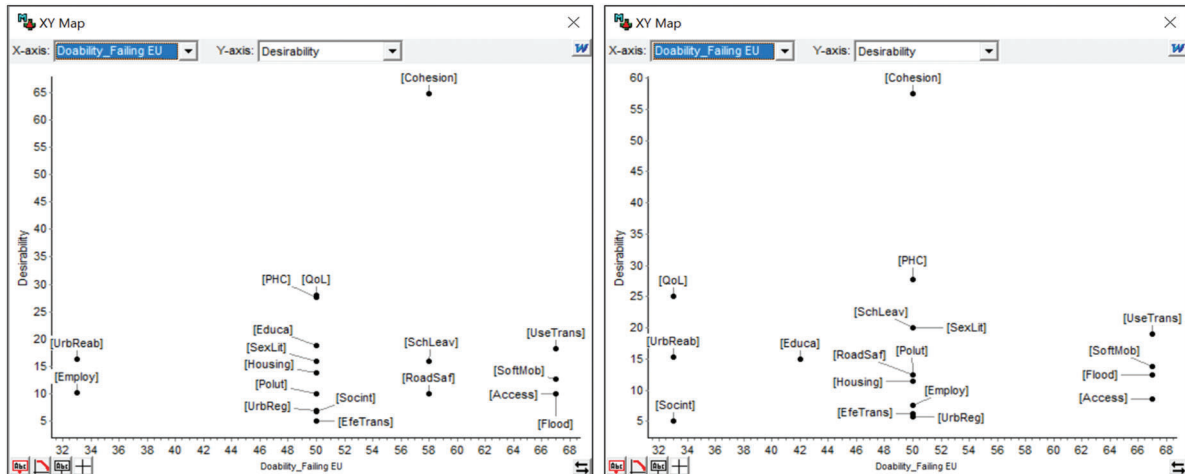


Fig. 10. Desirability–doability visual interactive graphs exemplifying the changes in policy positions under *Failing Europe* if a quick analytical procedure was followed (policy descriptions in Table 4).

3.5.2. Group sensitivity and robustness analyses

With respect to Challenge 5 (see Section 1), uncertainty was analyzed on several fronts with the group. It was first focused on the shared multicriteria desirability model. A revision of the critical situations, identified in the indicators of each objective, revealed potentially new ones in some policy units. Policymakers were then invited to design new measures to reinforce the “good” policy for the objective. Once again, this could not be done during the decision conference. It is worth noting that, in an *a posteriori* analysis, the resulting “better-than-good” new policy should receive an objective-specific score of desirability greater than 100, to keep the weights included in the built desirability model valid. A sensitivity analysis was performed on the weight of the [HS] objective, to address the divergent position that arose when setting the objective (see Section 3.2.2): It turned out that no significant change would occur in the results of the multicriteria desirability model, even if the weight of [HS] was equalized to the highest weight among the objectives. A robustness analysis was performed over the multicriteria desirability scores (shown in the overall thermometer of Fig. 8) given by Model (1) in face of effects on a policy performance, on one or several objectives, caused by external non-controllable conditions. It was found that a variation of at least ± 15 on the objective-specific desirability scores of the policies would be necessary to put the conclusion that policies [QoL] and [PCH] are more desirable than policies [Educa] and [UseTrans] in question. The top position of [Cohesion] in the final ranking was never compromised, which was not surprising in face of the bar chart at the right in Fig. 8. The positions of low desirability policies at the bottom are also robust, even if they are quite unstable among themselves.

The potential of using a $2 \times D$ graph for sensitivity analyses, on desirability and doability simultaneously, could also be tested during a decision conference. The interactive graph for the worst-case scenario is reproduced on the left in Fig. 10. The modified graph resulting from inputting equal weights for all objectives into the software when calculating the desirability scores and, simultaneously, the doability scores that would result from applying a simple majority rule to the initial individual judgments elicited under the worst-case scenario is displayed on the

right of the figure. The changes in the position of certain policies (e.g., the policy [Employ]) make clear that it is worth following an interactive process that includes eliciting judgments and their discussion toward group agreement rather than to follow a quick non-reflective pragmatic procedure.

4. Discussion

4.1. Discussing $2 \times D$ methodological options

This section focuses the discussion on the methodological options behind the proposed $2 \times D$ framework; Section 4.2 will analyze the extent to which *ex-post* feedback given by participants in the group process offers insights into the use and further development of $2 \times D$.

Let us start by further discussing why $2 \times D$ does not include scenarios in the protocols designed for eliciting desirability judgments. At first glance, this may be seen as an unrealistic assumption. An example is the case of policy [Polut], for which a drastic reduction in EU financial support would preclude the full implementation of most of its measures and consequently reduce the number of critical situations that [Polut] was expected to overcome – thus degrading the added value of the policy to improve the *SQ*. Of course, it is irrefutable that the consequences of implementing the policies will only occur in the future and are therefore inherently uncertain and, in turn, the policies risky. This may prompt other decision analysts to opt for a utility rather than value measurement approach (Keeney et al., 1993). $2 \times D$ takes a different methodological path and opts for modeling policy desirability assuming full implementation conditions to avoid desirability appraisal depending on doability levels. Note that this could happen in the Lisbon case where no policy was later judged as fully implementable. Uncertainty about policy consequences is dealt with in $2 \times D$ through extensive sensitivity and robustness analyses of the results obtained with the multicriteria desirability model meanwhile developed. Once again: “Theoretically, the choice mainly depends on the problem characteristics” (Keeney and von Winterfeldt, 2007, p. 236). Nonetheless, even in the presence of “deep” uncertainties (Karvetski and Lambert, 2012), a non-probabilistic multicriteria value model is a *practical* model (Keeney and von Winterfeldt, 2007): that is, “just good enough for the group to agree the way forward” (Phillips, 2007, p. 382), with the advantage that modeling value under certainty is technically simpler than modeling utility (Stewart, 2005) and consequently less cognitively demanding and less time-consuming.

Modeling the objective-specific desirability of consequences is appropriate and simpler as revealed by the practicability of the Smart Choices approach (Hammond et al., 1998). Nevertheless, $2 \times D$ diverges from Smart Choices after the step identifying the major uncertainties, for $2 \times D$ does not progress to quantifying them by assigning probability scores and calculating a measure of expected desirability. As referred in Section 1, the challenge of simplicity in modeling is central to the conceptualization of the $2 \times D$ framework. It is therefore relevant to avoid the well-known cognitive burden when making judgments of probability (Hogarth, 1975) because “deep” (non-probabilistic) uncertainties may well be present (Aven, 2013). $2 \times D$ adopts an alternative means of dealing with uncertainties, bounding them within ranges of plausible futures—by constructing two extreme scenarios (worst-case and best-case)—and analyzing the robustness of the results of the multicriteria desirability model. As remarked by Goodwin and Wright (2014), “scenario thinking

avoids any need to think probabilistically and allows a variety of viewpoints about the future to be reflected” (p. 409).

A second issue deserving discussion is related to the group interaction time needed to fully implement the $2 \times D$ framework. Modeling desirability by following the decompositional strategy of “divide and conquer” is time-consuming, and there are questions about whether a simpler and more expedited alternative holistic strategy might be followed to appraise the overall desirability of policies. However, this would preclude evaluating a policy in terms of individual contributions to improve the SQ on each fundamental objective, which is needed by policy managers for the allocation of limited public resources. Decision analysts should be aware that there is a price to pay (in process duration) to guarantee that time pressure will not give rise to the error of adopting procedures that lack theoretical significance and methodological rigour as is unfortunately often the case with *ad hoc* multicriteria models—an alert in line with Hammond et al. (1998). Finally, from the viewpoint of using $2 \times D$ in a real context of public health, such as the Lisbon case, we (inhabitants of the city) “personally don’t want some administrator to give two minutes of thought to the matter (...)”, paraphrasing Keeney (1992, p. 148). If time and other constraints affect the number of people involved, and geographic distance precludes the development of face-to-face group socio-technical processes, there are well-founded web-based alternatives, already used with good results as in the development of multicriteria PH indices in the framework of the EURO-HEALTHY project (Bana e Costa et al., 2022). Despite this, there are pros and cons to both types of interaction processes (see, e.g., Aubert et al., 2020, 2022) and which one is more appropriate—face-to-face or web-based or a combination of the two (see Vieira et al., 2020)—is context-dependent. That said, facilitators should be aware that adopting a decompositional procedure does not necessarily outweigh a holistic one, whatever the evaluation context (Morera and Budescu, 1998). In the $2 \times D$ framework, a holistic procedure is recommended each time there are questions of dependence; more specifically, on one hand, in the elicitation of desirability judgments and, on the other hand, in the evaluation of a policy’s doability, hereafter addressed separately as the third and fourth points of discussion.

The application of the MACBETH-based elicitation protocol for measuring the desirability of policies on each objective (see Table 1) implicitly assumes, as an elicitation working hypothesis, that judgments of difference of desirability between policies, for a fundamental objective, can be made *ceteris paribus*, that is, they are not affected by the policy performance levels on any other objective. It may be that the facilitator notices some difference dependency during the questioning process due to synergies between contributions to different objectives. The detected interdependent objectives should be merged to form a unique evaluation axis (see examples in Keeney, 1992; Bana e Costa and Beinat, 2005) for it to be possible to then apply the simple additive Model (1). Alternatively, a more complex aggregation model able to account for dependencies could be constructed (for theoretical details, see Dyer and Sarin, 1979); however, this would collide with the aim of simplicity, referred in Section 1 as a key to successful collaborative modeling.

It should be noted that, in general, two different paths can be followed to measure the relative desirability of an objective. In the first path, policies are directly pairwise compared to one another as in the protocol in Table 1. In the second path, policies are indirectly evaluated using a desirability function previously built upon the performance descriptor of the objective (as in Karvetski and Lambert, 2012). Each of these approaches has advantages and drawbacks, and both can be accommodated in $2 \times D$. In a context of spatial policies, with their performance levels varying across the

policy units of a geographical area, following the second path and constructing a separate desirability function for each objective would require the verification of additional assumptions. The difference of desirability between two policies in the territory could then be given by the average sum of their differences in desirability across the policy units. As far as we know, the only work that deals, from a methodological perspective, with public policies aided by MCDA when outcomes are spread over the territory is Simon et al. (2014). The difference toward $2 \times D$ is that, contrary to Simon, $2 \times D$ does not recommend following the decompositional path of assigning a desirability score to each policy on each spatial unit because this introduces complex problems of interdependence. Verifying such strong conditions seems unrealistic in most contexts, therefore also contradicting the $2 \times D$ core aim of simplicity in desirability modeling. Alternatively, the MACBETH voting protocol is used in $2 \times D$ to facilitate the elicitation of holistic judgments, by the direct comparison of policies taking all the critical situations that each policy is expected to resolve together.

The last issue of discussion relates to the $2 \times D$ adoption of a holistic path to appraise doability due to the unrealism of following a decompositional strategy to first appraise policy doability on each type of barrier separately. Indeed, the expected phenomena of interrelationships between the effects produced by the different types of barriers would preclude a simple additive aggregation of specific-barrier doability scores and would require the use of more complex aggregation procedures (as in the MACBETH-Choquet procedure; Oliveira et al., 2018).

4.2. Case insights for the framework

As noted in Section 3.1, *ex-post* feedback was obtained in individual semi-structured interviews with three selected policymakers, with similar questions asked of them. The following considerations are based on the notes taken by the interviewer. All three interviewees agreed that the framework was adequate for evaluating policies and found the way group qualitative judgments were elicited motivating. The $2 \times D$ visual graph was also recalled as providing a constructive perception and understanding of how different scenarios potentially affect the doability of policies. This is a key comment, as there were concerns when designing $2 \times D$ about whether the graph could possibly be perceived as complicated in practical settings. In the structuring phase, the interviewees also raised the need to move a step forward in defining the policies. It was recalled as doubtful whether considering the cross-effects of different policies in the multiple objectives would significantly affect the desirability of a policy in view that it might negatively affect the added value of another policy or that there might be synergies in jointly implementing a specific subset of policies. From a methodological viewpoint, these comments reveal, once again, the critical issue of interdependent policies. Simple structuring tools, such as the Analysis of Interconnected Decision Areas (Friend and Hickling, 1987, Luckman, 1967), are useful to form more horizontal packages of complementary and compatible actions/measures as, for example, in Bana e Costa et al. (2002).

Both the city councillor and the spatial planning chief officer stated that in order to bridge the evaluation and the operationalization of policies, the doability of each policy should be reassessed considering the following issues in more detail: (i) the doability judgments may fail to adequately consider governance and networking issues, which may easily change the expected outputs in the short term, namely, due to elections for new political decision-makers and changes in the relationships between the central and local governments and (ii) the deadlines for starting the

implementation of the selected policies frequently affect their outcome since circumstances normally change with time (as planning schedules and political timings seldom match). These comments are debatable in terms of the usefulness of a decompositional doability analysis versus the time required to build a complex doability aggregation model.

The public health expert believed that, from a public health management point of view, the policy evaluation phase should consider: (a) the different levels of implementation, as in many cases it is blind to geographical and administrative levels (e.g., policies that are applied to all civil parishes in the same way) and (b) the need for an explicit prospective analysis since the future implementation of selected policies affects the course of changes. In socio-technical terms, this is another source of uncertainty that can be addressed with sensitivity and robustness analyses.

5. Conclusions and future research

The $2 \times D$ framework advances knowledge on how to assist policymakers in evaluating and selecting spatial policies. $2 \times D$ is innovative in addressing, under an integrated and coherent methodological umbrella, all the challenges described in Section 1, for bridging the socio-technical gap between the support that a social decision-making process requires and what analytic techniques usually offer. Each part of the framework also involves innovative contributions. One cross-cutting contribution is the careful definition of interaction protocols, allowing the application of $2 \times D$ to be replicated in a specific context. The two-dimensional policy analysis, in terms of their desirability and doability combined with the modeling of uncertainties through scenario building, allows, on the one hand, the robustness of the desirability model results to be tested and, on the other hand, doability to be appraised under different futures that affect barriers to implementation. To our knowledge, this has not yet been reported in the literature. Last but not the least, the structuring of the problem of spatial policies by identifying critical situations across geographic units is also worth mentioning.

The framework's application proved adequate and robust in the Lisbon case. Specifically, the framework enabled the group of policymakers to (i) develop an understanding of the health problems that the city faces, (ii) appraise existing evidence—which integrates official documents—to tackle the identified problems, and (iii) balance the (multicriteria) desirability *versus* doability of policies in light of two contrasting scenarios. $2 \times D$ further follows well-established guidelines to conduct high-quality policy evaluation, namely, through the application of a context-specific and tailor-made approach. Policy evaluation was conducted transparently so that the participants involved could adhere to it in all phases. Value judgments were thus made explicit, and the use of a MACBETH technique allowed robust and sound elicitation protocols to be applied, which resulted in clear questions, conclusions, and points for attention, which was considered valuable and practical in the local policymaking context. The Lisbon case also provided insights regarding the role that scenario thinking can play in the appraisal and selection of policies. As stated in Grutters et al. (2015), “accepting uncertainty will not make the inevitable decisions in healthcare less painful, but will help to better allocate the scarce healthcare resources, and make these decisions more accountable and, therefore, acceptable” (p. 3).

Following the development and the successful application of the proposed $2 \times D$ framework, some aspects deserve further research. First, the role of scenarios in policy appraisal should be

studied in more detail. Despite the acknowledgment that doability can change in light of possible different futures, it was not fully explored how scenarios can affect the desirability of policies, or even further, if new policies should be designed to anticipate these scenarios. Nevertheless, at least one important conclusion can be drawn from the case: presenting policymakers with two clearly described extreme external scenarios and inviting them to think about policy doability separately under each scenario not only aligns the group and focuses their judgments but also establishes a quantitative doability range for each policy to mark out the setting of mitigation and elimination measures for barriers. Second, $2 \times D$ should be applied to other cases to test its adaptability to other regional and local contexts. This should be combined with research regarding other participatory processes, such as the use of web-based platforms engaging a higher number of participants (Vieira et al., 2020).

Acknowledgments

The authors would like to thank all stakeholders involved in the Lisbon real-world socio-technical process and the remaining consultation team members. The EURO-HEALTHY project (Shaping EUROpean policies to promote HEALTH equity) has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 643398. The authors received support from CEG-IST, FCT—Foundation for Science and Technology, I.P., under the project UIDB/00097/2020. The author Ana Vieira was also supported in her research by the Portuguese Foundation for Science and Technology under the MEDI-VALUE project (Grant no. PTDC/EGE-OGE/29699/2017).

References

- Ackermann, F., Eden, C., Rosenhead, J., Mingers, J., 2001. SODA-journey making and mapping in practice. In Mingers, J., Rosehea, J. (eds) *Rational Analysis in a Problematic World Revisited*. John Wiley & Sons, Chichester, pp. 43–61.
- Alvarenga, A., Bana e Costa, C.A., Borrell, C., Ferreira, P.L., Freitas, Â., Freitas, L., Oliveira, M.D., Rodrigues, T.C., Santana, P., Lopes Santos, M., Vieira, A.C.L., 2019. Scenarios for population health inequalities in 2030 in Europe: the EURO-HEALTHY project experience. *International Journal for Equity in Health* 18, 100.
- Angelis, A., Kanavos, P., 2017. Multiple criteria decision analysis (MCDA) for evaluating new medicines in health technology assessment and beyond: the advance value framework. *Social Science & Medicine* 188, 137–156.
- Aubert, A.H., Esculier, F., Lienert, J., 2020. Recommendations for online elicitation of swing weights from citizens in environmental decision-making. *Operations Research Perspectives* 7, 100156.
- Aubert, A.H., Lienert, J., Von Helversen, B., 2022. Gamified environmental multi-criteria decision analysis: information on objectives and range insensitivity bias. *International Transactions in Operational Research*. <https://doi.org/10.1111/itor.13206>
- Aven, T., 2013. On how to deal with deep uncertainties in a risk assessment and management context. *Risk Analysis* 33, 2082–2091.
- Bana e Costa, C.A., Beinat, E., 2005. *Model-Structuring in Public Decision-Aiding*. Operational Research Group, Department of Management, London School of Economics, London.
- Bana e Costa, C.A., Costa-Lobo, M.L., Ramos, I. A.J., Vansnick, J.C., 2002. Multicriteria approach for strategic town planning: The case of Barcelos. In Bouyssou, D., Jacquet-Lagrèze, E., Perny, P., Słowiński, R., Vanderpooten, D., Vincke, P., Bouyssou, D., Jacquet-Lagrèze, E., Perny, P., Słowiński, R., Vanderpooten, D., Vincke, P. (eds) *Aiding Decisions with Multiple Criteria: Essays in Honor of Bernard Roy*. Springer, Boston, MA, pp. 429–456.

- Bana e Costa, C.A., De Corte, J.-M., Vansnick, J.-C., 2012. MACBETH. *International Journal of Information Technology & Decision Making* 11, 359–387.
- Bana e Costa, C.A., De Corte, J.M., Vansnick, J.C., 2005. On the mathematical foundations of MACBETH. In Figueira, J., Greco, S., Ehrogott, M., Figueira, J., Greco, S., Ehrogott, M. (eds) *Multiple Criteria Decision Analysis: The State of the Art Surveys*. Springer, New York, NY, pp. 409–437.
- Bana e Costa, C.A., De Corte, J.M., Vansnick, J.C., 2017. M-MACBETH (beta) version 3.2.0: User's Guide. Available at: http://m-macbeth.com/wp-content/uploads/2017/10/M-MACBETH-Users-Guide_BETA.pdf (accessed: December 28, 2021).
- Bana e Costa, C.A., Ensslin, L., Cornêa, É.C., Vansnick, J.-C., 1999. Decision support systems in action: integrated application in a multicriteria decision aid process. *European Journal of Operational Research* 113, 315–335.
- Bana e Costa, C.A., Fernandes, T.G., Correia, P.V.D., 2006. Prioritisation of public investments in social infrastructures using multicriteria value analysis and decision conferencing: a case study. *International Transactions in Operational Research* 13, 279–297.
- Bana e Costa, C.A., Lourenço, J.C., Oliveira, M.D., Bana E Costa, J.C., 2014. A socio-technical approach for group decision support in public strategic planning: the Pernambuco PPA case. *Group Decision and Negotiation* 23, 5–29.
- Bana e Costa, C.A., Oliveira, M.D., Vieira, A.C.L., Freitas, L., Rodrigues, T.C., Bana E Costa, J., Freitas, Á., Santana, P., 2022. Collaborative development of composite indices from qualitative value judgements: The EURO-HEALTHY Population Health Index model. *European Journal of Operational Research* 305, 1, 475–492.
- Bana e Costa, C.A., Vansnick, J.-C., 1994. MACBETH—an interactive path towards the construction of cardinal value functions. *International Transactions in Operational Research* 1, 489–500.
- Baxter, R., 2015. *Operational Excellence Handbook: A Must Have for Those Embarking On a Journey of Transformation and Continuous Improvement*. Value Generation Partners, Naples, FL.
- Bertsch, V., Geldermann, J., Rentz, O., 2007. Preference sensitivity analyses for multi-attribute decision support. In Waldmann, K.H., Stocker, U.M. (eds) *Operations Research Proceedings 2006*. Springer, Berlin, pp. 411–416.
- Clemen, R.T., 1996. *Making Hard Decisions: An Introduction to Decision Analysis*. Brooks/Cole Publishing Company, Monterey, CA.
- De Corte, J., 2002. *Un logiciel d'Exploitation d'Information Préférentielles pour l'Aide à la Décision. Bases Mathématiques et Algorithmiques*. PhD thesis, University of Mons-Hainaut, Mons, Belgium.
- Dodgson, J.S., Spackman, M., Pearman, A., Phillips, L.D., 2009. *Multi-Criteria Analysis: A Manual*. London School of Economics and Political Science, Department of Economic History, London.
- Durbach, I.N., Stewart, T.J., 2012. Modelling uncertainty in multi-criteria decision analysis. *European Journal of Operational Research* 223, 1–14.
- Dyer, J.S., Sarin, R.K., 1979. Measurable multiattribute value functions. *Operations Research* 27, 810–822.
- Eden, C., 2004. Analyzing cognitive maps to help structure issues or problems. *European Journal of Operational Research* 159, 673–686.
- Fasolo, B., Bana e Costa, C.A., 2014. Tailoring value elicitation to decision makers' numeracy and fluency: expressing value judgments in numbers or words. *Omega* 44, 83–90.
- Fernandes, H.E., Ferreira, F.A., 2020. Health insurance risk assessment using cognitive mapping and multiple-criteria decision analysis. *International Transactions in Operational Research*. <https://doi.org/10.1111/itor.12895>
- Franco, L.A., Montibeller, G., 2010. Facilitated modelling in operational research. *European Journal of Operational Research* 205, 489–500.
- Freitas, A., Rodrigues, T.C., Santana, P., 2020. Assessing urban health inequities through a multidimensional and participatory framework: evidence from the EURO-HEALTHY project. *Journal of Urban Health* 97, 857–875.
- Friend, J., Hickling, A., 1987. *Planning under Pressure: The Strategic Choice Approach*, Butterworth-Heinemann, Oxford.
- Good, N., Ellis, K.A., Mancarella, P., 2017. Review and classification of barriers and enablers of demand response in the smart grid. *Renewable and Sustainable Energy Reviews* 72, 57–72.
- Goodwin, P., Wright, G., 2014. Scenario planning: an alternative way of dealing with uncertainty. In Goodwin, P., Wright, G. (eds) *Decision Analysis for Management Judgment*. John Wiley & Sons, Hoboken, NJ, pp. 387–422.
- Gregory, R., Failing, L., Harstone, M., Long, G., Mcdaniels, T., Ohlson, D., 2012. *Structured Decision Making: A Practical Guide to Environmental Management Choices*. John Wiley & Sons, Chichester.

- Grutters, J.P.C., Van Asselt, M.B.A., Chalkidou, K., Joore, M.A., 2015. Healthy decisions: towards uncertainty tolerance in healthcare policy. *PharmacoEconomics* 33, 1–4.
- Hammond, J.S., Keeney, R.L., Raiffa, H., 1998. The hidden traps in decision making. *Harvard Business Review* 76.
- Head, B.W., 2016. Toward more “evidence-informed” policy making? *Public Administration Review* 76, 472–484.
- Hogarth, R.M., 1975. Cognitive processes and the assessment of subjective probability distributions. *Journal of the American Statistical Association* 70, 271–289.
- Johnson, G., Scholes, K., Whittington, R., 2009. *Fundamentals of Strategy*. Pearson Education, Hoboken, NJ.
- Karvetski, C.W., Lambert, J.H., 2012. Evaluating deep uncertainties in strategic priority-setting with an application to facility energy investments. *Systems Engineering* 15, 483–493.
- Keeney, R.L., 1992. *Value-Focused Thinking: A Path to Creative Decisionmaking*. Harvard University Press, Cambridge, MA.
- Keeney, R.L., Raiffa, H., Meyer, R.F., 1993. *Decisions with Multiple Objectives: Preferences and Value Trade-Offs*. Cambridge University Press, Cambridge, MA.
- Keeney, R.L., Von Winterfeldt, D., 2007. Practical value models. In Edwards, W., Miles, R.F., Von Winterfeldt, D. (eds) *Advances in Decision Analysis: From Foundations to Applications*. Cambridge University Press, Cambridge, MA, pp. 232–252.
- Keller, L.R., Simon, J., 2019. Preference functions for spatial risk analysis. *Risk Analysis* 39, 244–256.
- Kirkwood, C., 1997. *Strategic Decision Making: Multiobjective Decision Analysis With Spreadsheets*. Duxbury Press, Belmont.
- Luckman, J., 1967. An approach to the management of design. *Journal of the Operational Research Society* 18, 345–358.
- Mateus, R.J.G., Bana e Costa, J.C., Matos, P.V., 2017. Supporting multicriteria group decisions with MACBETH tools: selection of sustainable brownfield redevelopment actions. *Group Decision and Negotiation* 26, 495–521.
- May, A.D., Page, M., Hull, A., 2008. Developing a set of decision-support tools for sustainable urban transport in the UK. *Transport Policy* 15, 328–340.
- Morera, O.F., Budescu, D.V., 1998. A psychometric analysis of the “divide and conquer” principle in multicriteria decision making. *Organizational Behavior and Human Decision Processes* 75, 187–206.
- Morestin, F., 2012. *A Framework for Analyzing Public Policies—Practical Guide*. Montreal, Québec, http://www.ncchpp.ca/docs/Guide_framework_analyzing_policies_En.pdf, (accessed: December 28, 2021).
- Oliveira, M.D., Lopes, D.F., Bana e Costa, C.A., 2018. Improving occupational health and safety risk evaluation through decision analysis. *International Transactions in Operational Research* 25, 375–403.
- Pardo Del Val, M., Martínez Fuentes, C., 2003. Resistance to change: a literature review and empirical study. *Management Decision* 41, 148–155.
- Parnell, G.S., Bresnick, T.A., Steven, T.N., Johnson, E.R., 2013. *Handbook of Decision Analysis*. John Wiley & Sons, Hoboken NJ.
- Pelissari, R., Oliveira, M.C., Abackerli, A.J., Ben-Amor, S., Assumpção, M.R.P., 2021. Techniques to model uncertain input data of multi-criteria decision-making problems: a literature review. *International Transactions in Operational Research* 28, 523–559.
- Phillips, L.D., 2014. Benefit-risk modeling of medicinal products: methods and applications. *Benefit-risk assessment in pharmaceutical research and development* 59–96.
- Phillips, L. D., 1984. A theory of requisite decision models. *Acta Psychologica* 56, 29–48.
- Phillips, L.D., 2007. Decision conferencing. In Edwards, W., Miles, R., Winterfeldt, V.D., Edwards, W., Miles, R., Winterfeldt, V.D. (eds) *Advances in Decision Analysis: From Foundations to Applications*. Cambridge University Press, Cambridge, pp. 375–399.
- Phillips, L.D., Bana e Costa, C.A., 2007. Transparent prioritisation, budgeting and resource allocation with multi-criteria decision analysis and decision conferencing. *Annals of Operations Research* 154, 51–68.
- Raiffa, H., 1968. *Decision Analysis: Introductory Lectures on Choices Under Uncertainty*. Addison-Wesley, Reading, MA.
- Rhorbach, B., 1969. Kreative nach regeln: methode 635, eine neue technik zum losen von problemen. *Absatzwirtschaft* 12, 73–75.
- Roy, B., 2010. Two conceptions of decision aiding. *International Journal of Multicriteria Decision Making* 1, 74–79.
- Santana, P., Freitas, Á., Stefanik, I., Costa, C., Oliveira, M., Rodrigues, T.C., Vieira, A., Ferreira, P.L., Borrell, C., Dimitroulopoulou, S., Rican, S., Mitsakou, C., Mari-Dell’olmo, M., Schweikart, J., Corman, D., Bana e Costa,

- C.A., 2020. Advancing tools to promote health equity across European Union regions: the EURO-HEALTHY project. *Health Research Policy and Systems* 18, 18.
- Schein, E., 1999. *Process Consultation Revisited: Building the Helping Relationship*. Addison-Wesley, Reading, MA.
- Shar, H., Compton, P., Anderson, M., Youngblood, A., 2011. Transportation model validation using Extreme-World method scenario construction. *Journal of the Transportation Research Forum*, 48, 105.
- Simon, J., Kirkwood, C.W., Keller, L.R., 2014. Decision analysis with geographically varying outcomes: preference models and illustrative applications. *Operations Research* 62, 182–194.
- Smith, J.E., Dyer, J.S., 2021. On (measurable) multiattribute value functions: an expository argument. *Decision Analysis* 18, 4, 247–256.
- Spaniol, M. J., Rowland, N. J., 2019. Defining scenario. *Futures & Foresight Science* 1, e3.
- Stewart, T.J., 2005. Dealing with uncertainties in MCDA. In Greco, S., Ehrgott, M., Figueira, J. (eds) *Multiple Criteria Decision Analysis: State of the Art Surveys. International Series in Operations Research and Management Science*. Springer, New York, pp. 445–466.
- Vieira, A.C., Oliveira, M.D., E Costa, C.A.B., 2020. Enhancing knowledge construction processes within multicriteria decision analysis: the Collaborative Value Modelling framework. *Omega* 94, 102047.
- Von Winterfeldt, D., Edwards, W., 1986. *Decision Analysis and Behavioral Research*. Cambridge University Press, Cambridge, MA.
- World Health Organization & Who Centre For Health Development (KOBE JAPAN), 2010. Urban HEART: Urban Health Equity Assessment and Response Tool. World Health Organization. Available at: <https://apps.who.int/iris/handle/10665/79060> (accessed: December 28, 2021).