Aubert, Alice H.; Esculier, Fabien; Lienert, Judit

Article

Recommendations for online elicitation of swing weights from citizens in environmental decision-making

Operations Research Perspectives

Provided in Cooperation with:
Elsevier

Suggested Citation: Aubert, Alice H.; Esculier, Fabien; Lienert, Judit (2020) : Recommendations for online elicitation of swing weights from citizens in environmental decision-making, Operations Research Perspectives, ISSN 2214-7160, Elsevier, Amsterdam, Vol. 7, pp. 1-13, https://doi.org/10.1016/j.orp.2020.100156

Operations Research Perspectives 7 (2020) 100156

Contents lists available at ScienceDirect

Operations Research Perspectives

journal homepage: www.elsevier.com/locate/orp

Recommendations for online elicitation of swing weights from citizens in environmental decision-making

Alice H. Aubert[a,*], Fabien Esculier[b,c], Judit Lienert[a]

[a] Eawag: Swiss Federal Institute of Aquatic Science and Technology, Überlandstrasse 133, Dübendorf CH-8600, Switzerland
[b] Laboratoire Eau, Environnement et Systèmes Urbains (LEESU), AgroParisTech, École des Ponts ParisTech (ENPC), Université Paris-Est Marne-la-Vallée (UPEMLV), Université Paris-Est Créteil Val-de-Marne (UPEC): UMR MA-102, LEESU, ENPC, 6-8 avenue Blaise Pascal, Champs sur Marne cedex 2 77455, France
[c] Milieux Environnementaux, Transferts et Interactions dans les hydrosystèmes et les Sols (METIS), Centre National de la Recherche Scientifique (CNRS), École Pratique des Hautes Études (EPHE), Sorbonne Université: UMR7619, METIS, UPMC, Case courrier 105, 4 place Jussieu, Paris 75005, France

ARTICLE INFO

Keywords:
Behavioural Operational Research
OR in environment and climate change
Learning
Multi-Criteria Decision Analysis
Decision Support System
Public participation
E-democracy

ABSTRACT

There is a growing demand for public participation in environmental decision-making. However, it is unclear how a large number of citizens can best engage in such complex public policy decision processes. This need from the civil society challenges the OR community to develop online decision-making tools. This article reports on a feasibility assessment of swing weight elicitation, implemented online, for real-world decisions about future wastewater infrastructure. Eliciting weights with the swing method is common in MAVT/MAUT, but not online. A total of 298 affected citizens from the Paris region answered the online swing weight elicitation survey. Another 357 citizens directly rated objectives. Three aspects of learning in the context of MCDA were considered: did participants learn facts about the wastewater topic? Did they comply with the swing elicitation process, i.e. follow the instructions? Did participants learn about their preferences? Factual learning was limited. Process compliance was really low (12%), leading to a number of recommendations for improving the interface for online swing weight elicitation. The collected preferences differed statistically significantly between the compliant and non-compliant participants, and also between the non-compliant and direct rating respondents. This emphasised the effect of the elicitation method on preference construction. Moreover, more participants experienced a strengthening of pre-existing opinions than a change in opinion, and most reported being uncertain about their answers. This calls for better understanding process learning and preference construction. We discuss our developed procedure for online swing weight elicitation, recommend ways to improve swing online surveys, and suggest interesting future research lines that would allow empirically verifying our propositions.

1. Introduction

1.1. Environmental MCDA needs to be more participatory

The demand for public participation in public policy in general [1,2], and in environmental decision-making in particular, is growing [3–8]. Citizens want their opinions to affect decision-making, as opposed to being solely informed [9]. For some academics, social participation is essential for the success of sustainability policies [10]. The literature discusses many reasons to justify increased public engagement, from instrumental to normative rationales [11]. One argument is that including more – informed – people (public and stakeholders) improves the outcomes of the decision-making process [5]. In particular, citizen engagement broadens the range of opinions so that potential conflict lines become more obvious [12,13]. User involvement, or citizen engagement, can also increase public acceptance of the decision [7,14], and thus facilitate implementation, as it contributes for instance to legitimize an innovative sustainable alternative. Other presumed contributions are found in the comprehensive recent review by Baker and Chapin [11] and are listed hereafter. Participation can increase efficiency of decisions (e.g. cost effectiveness), promote collective action, adaptive capacity, and community development, in particular through social learning. It can also deepen democracy, and, in some cases, it fulfils agency mandates [11]. Obviously, participation is not a remedy for "all" public policy and environmental decision issues [5], and it may come with additional costs [15]. However, there are many cases where stronger public engagement is highly justified. This call for increased participation is challenging the decision analysis and environmental management communities, who need to adapt their practices to meet the demand [16–18]. In other words, formal decision-

* Corresponding author.
E-mail address: alice.aubert@eawag.ch (A.H. Aubert).

https://doi.org/10.1016/j.orp.2020.100156
Received 29 November 2019; Received in revised form 19 June 2020; Accepted 20 June 2020
Available online 25 June 2020

making methods should be adapted to meet the specific requirements of engaging a large number of people.

In this article, we focused on such an adaptation of the structured participatory process of Multi-Criteria Decision Analysis (see Introduction to MCDA, Section 1.2). Specifically, we aimed to collect preferences – in terms of the relative importance given to objectives (Introduction Section 1.3) – from many affected citizens. Concurrently, we evaluated the feasibility of the developed online survey, based on three aspects of individual learning in MCDA (Introduction Section 1.4).

1.2. Some MCDA (MAVT) theoretical inputs

Multi-Criteria Decision Analysis (MCDA) and Value-Focused Thinking (VFT) are commonly used methods for complex environmental decision-making [19–21]. They are a transparent way to deal with issues involving difficult trade-offs and potentially conflicting interests [e.g. [22] and references therein]. The problem is framed, and structured into a set of objectives j. Various alternatives i are identified and how they perform for each of the objectives is estimated (a_{ij}). Then, subjective preferences are elicited. This enables to aggregate the performance of each alternative for each objective a_{ij}. These subjective preferences include, for instance in Multi-Attribute Value Theory (MAVT), the marginal value functions for each objective v_j, the weights given to each objective w_j, and the aggregation model. The marginal value functions v_j transform the attribute levels measuring the fulfilment of objectives (or prediction a) of alternative i for objective j, a_{ij}, (measured with a unit, e.g. Euro for "low costs") to an interval scale (e.g. 0 to 1). The weights w of objective j, w_j, are scaling constants that represent the relative importance of one objective relative to the others. Last, given that the three axioms of simple and mutual preferential independence and difference independence (or compensation) are met [23], the additive aggregation model can be used to calculate the overall value of alternative i, $v(a_i)$ [24,25]. Formally, the MAVT additive model is:

$$v(a_i) = \sum_{j=1}^{m} w_j . v_j(a_{ij}) \quad (1)$$

Eliciting the subjective preferences, including the weights, is one of the main participatory parts of MCDA [23,25–27]. Face-to-face interviews or group workshops are the most common means to collect the preferences of the stakeholders, and have been applied for many environmental decisions [19]. However, these direct interactions are time-consuming and constrain the number of participants. To meet the societal demand for more participation [5], MCDA practice has to evolve. Online tools and civic tech for e-democracy allow a broad public to participate in decision-making [9,28]. These tools and technologies make decision-making accessible to many and speed up the elicitation [29].

1.3. Online swing weight elicitation needs real world testing

In the following, we focused on weight (w_j) elicitation, because it is a "crucial" step in MCDA that can strongly affect the results. It is cognitively demanding, prone to biases, and to some extent method dependant [26]. More specifically, we focused on the swing weight elicitation method, which belongs to the ratio weights procedures [26], used in MAVT. The attribute weights w_j are derived by normalising the sum of given points so that the sum of weights equals one. Swing weights describe the relative importance of the improvement of an objective's attribute from its worst possible to best possible level, compared to the improvement from worst to best of the other objectives' attributes [23]. The major advantage of swing weight elicitation, compared to the trade-off method, is that it does not require knowing the shape of the marginal value functions v_j [23]. Others reported that swing elicitation produced more stable weights compared to the smart/swing variant [30]. In addition, when used in a hierarchical fashion, swing weight elicitation minimises the occurrence of the equalising bias [31].

Online swing weight elicitation has been experimentally tested in the lab [29,32]. Authors suggest that "research should focus more on the practical applications that tend to be more complex than simple test tasks used in experiments" [29]. Others warn that unassisted weighting, with a questionnaire, possibly online, may elicit weights that wrongly reflect the opinion of the respondent [33]. To our knowledge, only one real case application of unassisted online weight elicitation for complex environmental issues exists [30]. Authors find it promising, but state that more evidence is needed. Recently, there is a general call to increase research on environmental decision-making processes in real-world interventions [6,34–36]. Consequently, our study aimed at testing the feasibility, i.e. the practicality and understandability, of online elicitation of swing weights in a real-world application, from affected citizens. Our study contributes to establishing good practice guidelines.

1.4. Assessing weight elicitation tools

Many consider MCDA [10,22,33,37], or other operations research tools and methods [38–41], as learning processes. In the context of preference elicitation, when one considers that preferences are constructed [42], learning is implicit. Building on this literature and the transformational learning theory for individuals [43], a framework for learning in MCDA was proposed [18]. It considers preference construction as a reflection on contents [43], composed of two aspects: *factual learning* and *preference learning*. Factual learning consists of discovering, understanding, and acquiring data about the system or decision problem at stake. It contributes to reinforcing or changing the internal representation of the problem. Preference learning consists of valuing the various aspects of the problem. It contributes to strengthening or changing the preferences. In addition, Mezirow's process reflection [43] is interpreted in the context of preference construction as learning about the tool, and/or method of elicitation, if not more generally about a way to make a decision.

Thus, assessing the feasibility, i.e. practicality and understandability, of the online weight elicitation tool requires assessing factual learning, preference learning and process learning. This led to our three main research questions (RQ):

RQ1. Did participants learn *facts* about wastewater management during the proposed online survey to elicit swing weights?

RQ2. Did participants follow the *process* instructions of the online survey to elicit swing weights?

RQ3. Did participants learn about their *preferences* during the online survey to elicit swing weights?

Another, more exploratory, research question emerged from a preliminary analysis of RQ2. The results indicated that very few participants actually complied with the instructions. Because we had recruited the participants through a panel survey company, we suspected that they were accustomed to filling in marketing surveys, which mostly consist of direct rating (according to the co-authors experience). Thus, we formulated the fourth RQ as follows:

RQ4. Did the participants who did not comply with the process to elicit swing weights perform a direct rating of objectives?

Section 2 of the paper introduces the methods. It includes a presentation of the online swing questionnaire, a summary of the case study, the measures used for the assessment, and information about data analysis. Section 3 reveals the results of the assessment. We discuss results in Section 4: online survey to elicit swing weights is feasible, but requires further research and development to design an improved interface that reduces the encountered problems. Our conclusions are summarized in Section 5.

2. Materials and method

2.1. Design of the online interface to elicit swing weights

We summarize the purposes of the interface hereafter. First, the interface should enable to elicit the weights from the respondents (i.e. facilitate preference learning), and save the elicited weights in a database. Second, in order to meet the prior aim, because the targeted audience is laypersons, the interface should inform the laypersons about the main decision issue, and the objectives to consider when deciding how to address the main issue (i.e. facilitate factual learning). Third, it should be compliant with the swing weight elicitation process, as described in the textbooks e.g. [23] (i.e. facilitate process learning), and follow the recommendations from the existing building code [44]. Fourth, it should improve a previous attempt by Lienert, Duygan and Zheng [30]. They had faced the following inconsistencies: rating (i.e. assigning weights) inconsistent with ranking of objectives, and rating of the worst-case hypothetical alternative different from zero (where all objectives are on their worst level and which should be zero according to swing convention). They suggested using the inputs of the ranking step to present the hypothetical alternatives in the right order for the following rating step. In addition, we list some constraints that we were facing. We had to develop the tool using a survey platform (no resources for a stand-alone software). We chose the Qualtrics platform (www.qualtrics.com, retrieved on 19.07.2018), as it was used for the previous attempt [30]. The following part describes the survey in more detail. For screenshots, please see the supplementary information (Section SI 2a).

First, we introduced the aim of the survey, and emphasised why the decision at stake was important and relevant for the participants. This real decision concerning the wastewater infrastructure for Paris is described in a separate paper [45]. The description of the first set of objectives, namely the objectives of the first branch of the objectives hierarchy, followed. We consistently presented the following elements for each objective: the objective name, its definition and an explanation of why the objective is important, its attribute (in which dimension it is estimated), the status quo in Paris today, the worst and best possible predictions for the alternatives that are considered and what the expected impact for these values are. An English version of the survey text is presented in the SI, and one example is given in Fig. 1. This informative part targeted factual learning.

Second, the weight elicitation part followed (see screenshots in SI 2 and Fig. 2). As first step of weight elicitation, the dominated worst-case hypothetical alternative was described (all attributes of the objectives to be compared are at their worst level), and represented graphically. A small image presented all objectives names, their measurement unit, and their worst-case prediction (clearly signalled by an unhappy red emoticon) in a table-like format. Hereafter, we refer to these table-like images representing hypothetical alternatives as vignettes. Thereafter, we introduced the swing hypothetical alternatives (with only one objective at the best level, and all others on their worst), again represented with vignettes. They depicted in green the objective improved to its best level, with a green happy smiley near the best-case prediction achieved by this objective. This first step is an information step. Second, the instructions asked the participants to order the vignettes representing the hypothetical alternatives by dragging and dropping in order of preference from the most preferred (at the top) to the least preferred (at the bottom). The wording of the instructions emphasized that they were prioritizing the improvement of an objective's attribute from its worst possible to best possible level, compared to the improvement from worst to best of the other objectives' attributes. When the ordering of the vignettes representing the hypothetical alternatives reflected their preference, the participants clicked next to move on to the next page. In the third step of weight elicitation, all vignettes representing the hypothetical alternatives were presented to the participants in the order of their preference, including the worst-case hypothetical alternative at the bottom. Each vignette was connected to a slider. The most preferred hypothetical alternative at the top received 100 points by default, the worst-case alternative at the bottom 0 points. The participants were asked to rate the hypothetical alternative(s) in between, by adjusting the slider according to their preference. The instructions explained that the rating had to be relative (translation from originally French questionnaire): "Please consider the relative importance. For instance, if you like a hypothetical alternative half as much as another, it should have half the rating. You can give equal rating if the hypothetical alternatives are equally good. As they are ordered according to your preference, the top ones should get a higher (or equal) rating."

This swing process, consisting of (1) the informative part about the objectives, (2) rank-ordering of vignettes representing the hypothetical alternatives, and (3) relative rating, was repeated for each of the four branches of the objectives hierarchy Fig. 3, branches B1 to B4). To obtain the weights of the four upper level objectives, there was a fifth swing process. This fifth swing process (later referred to as "Up", for upper level) considered the most important objective of each of the four branches, i.e. one objective coming from each branch, specifically the one with the highest weight within that branch. Participants were asked a fifth time to (1) rank-order the vignettes representing the hypothetical alternatives, and (2) rate them.

Practically, this required coding all the 24 possible combinations (for an example, see SI 2b). Three colleagues carefully pre-tested the usability of the survey, the wording of the instructions, and the coding of all 24 combinations of objectives. Their feedback was integrated into the tested version. Four native speakers proof-read the French translation of the survey text.

For comparative assessment, we developed an online survey with direct rating of the objectives. This online survey mimicked the current practice of non-MCDA surveys by asking for a direct rating of the objectives. There, the identical textual information about an objective was presented, namely: the objective name, its definition and an explanation of why the objective is important, its attribute (in which dimension it is estimated), the status quo in Paris today, the worst and best possible predictions for the alternatives that are considered, and what the expected impact for these values are. A slider allowing giving "importance points" from 0 to 100 directly followed. The participants directly rated the lower level objectives. A single webpage displayed all objectives of one branch only. This step-wise focus on only one branch could have emphasized relative rating by comparing only the objectives within this branch, although the instructions did not emphasize this. We were able to use direct rating for comparison because it used the same evaluation scale as swing (0 to 100).

	The five management alternatives (management 1–5) can also be evaluated according to how well they increase societal welfare.
Hea	For instance, it [a management alternative] should offer **high possibility of swimming in rivers**. Wastewater contains pathogens, such as some strains of *E. coli*, which can cause, for example, diarrhea, nausea or vomiting. People may get in contact with contaminated water that reaches rivers from the wastewater system. The *E. coli* discharge into the river is estimated by the number of *E. coli* colonies that appear in a standard dish of 100 ml of river water. According to European rules, swimming in rivers is allowed if there are less than 900 *E. coli* colonies in 100 ml of river water. **Currently**, we estimate that 17'100 colonies of *E. coli* appear in 100 ml of water from the river Seine. This is the **worst case**. It means that swimming in the Seine is not allowed because the health risk is high. In the **best case**, *E. coli* in the river Seine can be reduced to 8 colonies in 100 ml of river water. In this best case, there is low health risk and swimming in the Seine would be allowed.

Fig. 1. Example of description of the objective High possibility of swimming in rivers". This sub-objective belongs to the higher-level objective "High societal well-being". The description of each objective was presented to the participants at the beginning of the survey. Note, the original survey language (and thus text) was French.

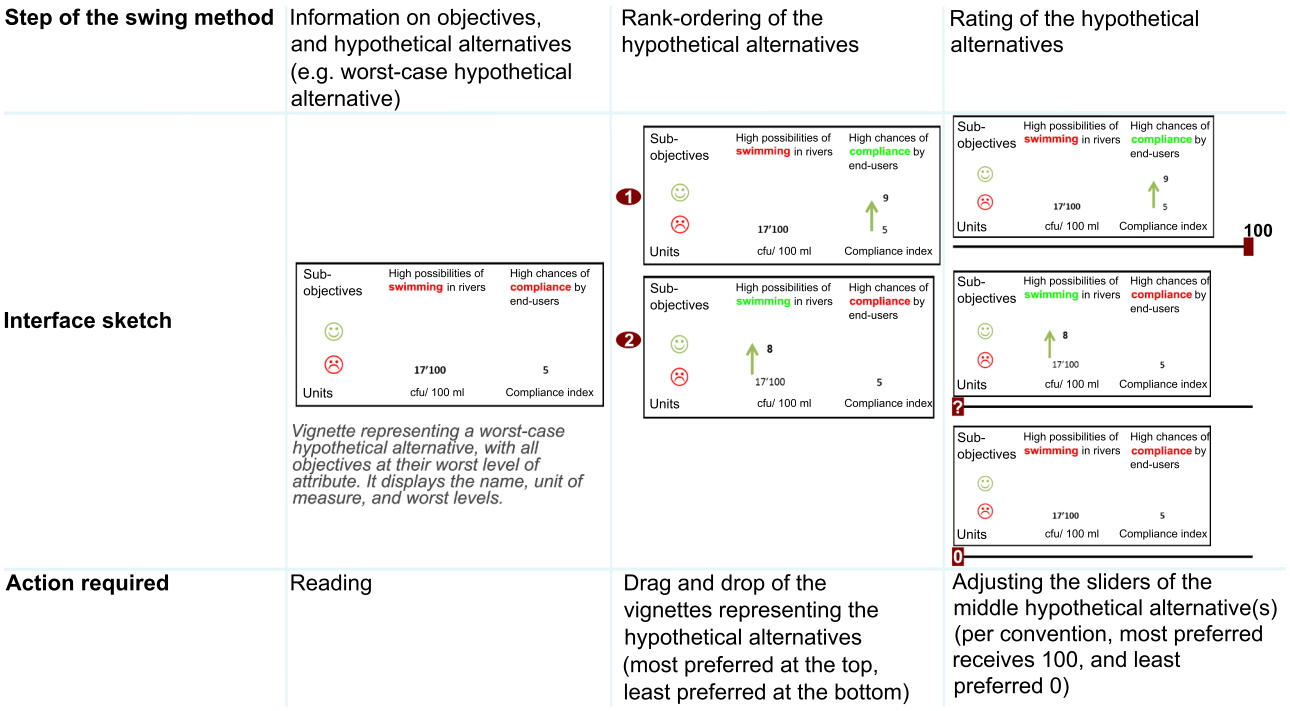

Fig. 2. Swing weight elicitation. The three steps of swing weight elicitation and the corresponding sketches of how the interface presented the information, and required actions from the participants. Screenshots of the interface are available in SI 2.

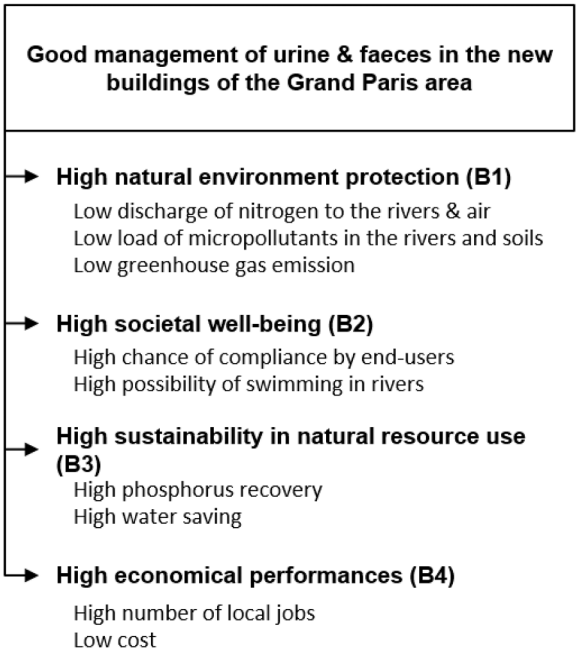

Fig. 3. The hierarchy of objectives for the Paris region case study on wastewater infrastructure. The relative importance (weights) of the nine sub-objectives was elicited from the citizens.

2.2. Summary of the case study

The authority responsible for wastewater management in Paris region is challenged to maintain a good management of urine and faeces in the Grand Paris region in the future. Due to continuous population growth and climate change, the current central wastewater treatment plant is reaching its capacity limits. We conducted a stakeholder analysis and invited representatives of all parties to a problem-structuring workshop in July 2016. Representatives of all parties (e.g. wastewater engineers, city planners, authority representatives) joined to define the objectives and generate alternatives. Only agriculture and health parties were missing, as they declined the invitation. Another paper focuses on the workshop and its outcomes [45]. After the workshop, we focused on the wastewater authority's need: learning about what the affected citizens value with help of online surveys, in order to make an acceptable –and likely not contested– decision. To find out which alternative (s) best meet the preferences of the citizens, we elicited the relative importance of objectives, i.e. the weights given to objectives. This focus on objectives is at the core of Value-Focused Thinking [25]. We aimed to avoid decision-making based on heuristics, which typically occurs when directly judging alternatives in a very complex case. We used the two-level and four-branch objectives hierarchy presented in Fig. 3 (see SI 1a for the description of the objectives). We used the elicited weights to evaluate five alternatives, considered for new building areas around Paris, namely a status quo alternative (centralised system as it is today, alt.1), two urine source separation alternatives: one concentrated urine (alt.2) and the other stored urine (alt.3), a vacuum system alternative (alt.4), and a composting alternative (alt.5) (see SI 1b for complete description of alternatives and SI 1c for the prediction matrix; more details in [45]).

2.3. Participants

The decentralized alternatives to the status quo require different toilets and a local processing of the wastewater (sometimes in the buildings where people live and work). Thus, the decision on the future wastewater management affects the inhabitants of the region, both directly (as some alternatives will have effects on their daily life, e.g. using different types of toilets) and indirectly (as the taxes they pay fund the infrastructure).

In December 2017, 542 inhabitants from the region Île-de-France were invited to answer the online survey to elicit swing weights by the panel company MadeInSurvey (selected after a call) (fr.madeinsurveys.com, retrieved on 19.07.2018), with quotas on gender, age (from 18 to 74 years old), and socio-professional category in order to represent the population of the region (see SI 3a-b). We used quotas to

reduce the selection bias due to the panel of the survey company. The participants received a monetary incentive, according to MadeInSurvey policy.

In December 2017, the same panel company invited to the online survey with direct rating of objectives 544 inhabitants from the region Île-de-France, following the same socio-demographic quotas as for the online survey to elicit swing weights (SI 3c-d). No participant was invited to both surveys. In total, 1′086 inhabitants from the Paris region were invited.

2.4. Measures used for the assessment

RQ1. Did participants learn facts about wastewater management during the proposed online survey to elicit swing weights? Factual learning was measured thanks to a knowledge test answered before and after the weight elicitation. There was one question per objective, some being multiple choice questions [30] (SI 5a). The score summed the number of correct answers: it varied from zero (no correct answer) to 18 (all correct answers were selected). We assumed that factual learning occurs if the final knowledge score is higher than the initial knowledge score.

RQ2. Did participants follow the process instructions of the online survey to elicit swing weights? We used process compliance as a proxy for process learning. Process compliance was measured by the fulfilment of the three very explicit instructions, at each of the five steps of the swing weight elicitation survey. These steps were: B1-B4 for the sub-objectives of each of four branches of the hierarchy, and one step to elicit the weights of the four upper-level objectives across branches (Up, Fig. 3). The following three measures were used: Did participants indeed rate the most preferred hypothetical alternative (where most-important objective is on its best level, all others on their worst level) with 100 points? Did they rate the worst-case hypothetical alternative (all objectives are on their worst level) with 0 point? Was the rating of the hypothetical alternatives (i.e. numbers between 100 and 0) in accordance with the previously established rank-ordering of preference (i.e. 1 = most important objective to improve, 2 = second most important … etc.)? These measures were 0/1 scales (not compliant/compliant). Process compliance occurs if participants correctly follow these instructions of the process.

RQ3. Did participants learn about their preferences during the online survey to elicit swing weights? Given the lack of established protocols to measure preference learning [46], we answered this question using three feedback questions. These questions were inspired from literature on learning in MCDA [18], and the proposition that participants give greater confidence to constructed preferences [46]. Two were yes-no questions. The first asked whether a change of opinion regarding some of the objectives had occurred, and the second whether the opinion regarding some of the objectives was strengthened. The third question asked how certain the participants were about the answers they provided.

RQ4. Did the participants who did not comply with the process to elicit swing weights perform a direct rating of objectives? We compared the structure of the weight distribution, based on the spread of weights and the relationship of the mean weight received by the objective as a function of its rank. These measures are explained hereafter.

Given the results on process compliance (RQ2), we split the sample into two to calculate the weights. One sub-sample, termed “swing”, contained the answers of those that complied with the concept of the worst-case hypothetical alternative: it contained the answers of those who gave this dominated alternative the lowest rating. The other “invalid swing” sub-sample gathered those answers that did not follow this instruction, i.e. the worst-case hypothetical alternative (where all objectives are on their worst level) received a rating higher than at least one of the other hypothetical alternatives with one objective at the best level.

The weights for the swing sub-sample were calculated as follows. First, we calculated local weights, i.e. normalised the ratings within each branch by the sum of the ratings given to the hypothetical alternatives of the branch (so that within each branch, the sum of weights equals 1). Second, we normalised the ratings at the upper level of the hierarchy by the sum of ratings given to the hypothetical alternatives of the upper level (so that at the upper level, the sum of weights equals 1). Third, we calculated the global weights (so that the sum of weights of the nine sub-objectives equals 1). This is formulated in Eq. (2).

$$w_r = \frac{p_r}{\sum_{i=1}^{m} p_i} \cdot \frac{p_R}{\sum_{j=1}^{n} p_j} \tag{2}$$

with w_r the global weight of objective r, p_r the rating given to the objective r (lower level in the hierarchy), m the number of objectives within a branch, p_R the rating given to the objective R (corresponding to the upper level in the hierarchy), n the number of objectives at the upper level (equals the number of branches).

Weights for invalid swing were calculated as follows: we normalised the ratings given to the hypothetical alternatives by the sum of the nine ratings given to the hypothetical alternatives, so that the sum of weights of the nine sub-objectives equals 1 (Eq.3).

$$w_r = \frac{p_r}{\sum_{k=1}^{o} p_k} \tag{3}$$

with w_r the global weight of objective r, p_r the rating given to the objective r, o the number of lower level objectives.

We calculated the weights from the direct rating survey following Eq. (3), normalising the rating given to one objective by the sum of all the ratings.

Thereafter, we compared the invalid swing weight distribution with the weight distribution from swing, and direct rating. First, we calculated the spread of weights, as the difference between the highest and lowest weights. The spread of weight varies between zero (all weights are equal) and one. We observed the proportion of participants who gave equal weights (the spread of weights is zero), and the proportion of participants whose spread of weights is above the threshold of 0.11 (this represents the weight received by an objective in a nine-objective problem if the weights are equally distributed [47]). Second, we observed the relationship of the mean weight received by the objective as a function of its rank, as done in [47]. In particular, we calculated nine independent t-tests to compare weights of each rank; we described the shape of the non-linear curve representing the average weights as a function of the rank; and we performed an analysis of variance with the weights as dependant variable, the sample group was used as a between-participant factor, and the rank as a within-participant factor. Participants from the invalid swing sub-sample were considered to have performed a direct rating if the above-mentioned criteria describing the weight distribution differed statistically significantly from the swing weight distribution, and did not differ from the direct rating survey. Table 1 summarises the research questions of this paper.

2.5. Data analysis

The statistical analyses were performed using R project for statistical computing [48]. The tests used for each research questions are described in the specific Results Section 3. They were performed with alpha level of 0.05 and beta 0.95.

In addition, we controlled that the actual time needed to answer the survey did not explain the number of unfollowed instructions (visualizing the data, and performing a correlation analysis). Because this was not the case, we were able to set a cut-off line of 15 min as minimum required time (SI 4a). No respondents from the valid swing sub-sample were faster than 15 min. It is also the reasonable time needed to read all the text (description of the objectives, and the instructions).

Table 1
Summary of the research questions (RQ), and the data used to answer them.

	Research questions	Assessment	Data considered
RQ1	Did participants learn facts about wastewater management during the proposed online survey to elicit swing weights?	Factual learning occurs if the final knowledge score is higher than the initial knowledge score.	Swing survey (N = 298)
RQ2	Did participants follow the process instructions of the online survey to elicit swing weights?	Process compliance occurs if participants correctly followed the three instructions of the process.	Swing survey (N = 298)
RQ3	Did participants learn about their preferences during the online survey to elicit swing weights?	Preference learning occurs if the three self-reported answers indicated that this was the case.	Swing survey (N = 298)
RQ4	Did the participants who did not comply with the process to elicit swing weights perform a direct rating of objectives?	Participants were considered to have performed a direct rating if (1) the weight distribution statistically significantly differed from the one of swing and (2) if there was no statistically significant difference with direct rating.	1. Sub-sample "invalid swing" of the swing survey (N = 262) 2. Sub-sample "swing" of the swing survey (N = 36) 3. Direct rating survey (N = 357)

3. Results

3.1. Description of the population

In December 2017, 542 inhabitants from the Paris region were invited to answer the online survey to elicit swing weights. The completion rate was 68% (number of completed answers: 365). This is comparable with previous studies (e.g. 64% in Lienert, Duygan and Zheng [30]). Of these, 67 respondents (18%) were removed for answering the survey too quickly (less than 15 min; see Section 2.5). The gender, age and socio-demographic distributions of the remaining 298 answers were representative of the regional statistics: none of the Pearson's Chi squared test showed significant differences in the distributions between our population sample and the official statistics (SI 3a and SI 3b).

Also in December 2017, 544 inhabitants from the Paris region were invited to answer the online direct rating survey, used for research question 4. The completion rate was 79% (number of completed answers: 431). Of these, for consistency with the swing survey, we removed the 17% faster respondents (i.e. the 74 respondents who answered in less than 11 min). This was justified by the facts that (1) the descriptions of objectives were the same lengthy text to read, but (2) there were fewer questions, and the questions were easier. The gender, age and socio-demographic distributions of the remaining 357 answers were representative of the regional statistics (analysed with Pearson's Chi squared test as above SI 3a and SI 3c). The company providing access to the panel guaranteed that no respondent answered both surveys. The population samples of the swing and direct rating surveys are similar in terms of gender, age, and socio-demographic distributions (SI 3d).

3.2. Factual learning

RQ1: Did participants learn facts about wastewater management during the online survey to elicit swing weights? According to the Wilcoxon statistical test for paired samples (Table 2), the hypothesis H0 (final and initial knowledge scores of the swing survey follow the same distributions) was rejected. The final knowledge score was statistically significantly higher than the initial knowledge score (about one point of 18 possible points). Respondents who answered the online survey to elicit swing weights did learn about wastewater management. However, this learning is limited.

3.3. Process compliance

RQ2: Did participants follow the process instructions of the online survey to elicit swing weights? During the first swing elicitation (step B1 for objectives of branch "High natural environment protection"; Fig. 3), at best half of the participants (50.3%) followed the instruction that "the most preferred hypothetical alternative receives 100 points" (Fig. 4). However, as worst case, only 23.5% followed the instruction "the worst-case hypothetical alternative receives 0 point" in step B1. The third instruction was followed by 38.9% of participants, namely that "the rating of the hypothetical alternatives (i.e. assign numbers between 100 and 0) should correspond to the previously established order of preference (i.e. 1 = most important objective, 2 = second most important … etc.).

In the following, we compared the percentage of participants who followed each instruction at the first swing weight elicitation step (assign weights to the sub-objectives within branch B1) and the last step (Up; assign weights for upper-level objectives across all four branches). At the fifth swing (Up; upper level objectives), 9.7% more participants followed the instruction that "the most preferred hypothetical alternative should receive 100 points" (Fig. 4, plain line) than at the first B1 step. This instruction implies that the participant did not modify the default value set to 100 for the most preferred hypothetical alternative. Equally, 9.7% more participants followed the instruction that "the worst-case hypothetical alternative should receive 0 point" at the last step compared to the first step (Fig. 4, long dashed line). This instruction implies that the participant did not modify the default value set to 0 for the worst-case hypothetical alternative. Note that these 9.7% participants are not systematically the same as for the instruction

Table 2
Factual learning.

	Swing (total, N = 298)					Swing "valid" (N = 36)					Swing "invalid" (N = 262)				
	Min	Med	Mean	Max	Paired Wilcoxon test	Min	Med	Mean	Max	Paired Wilcoxon test	Min	Med	Mean	Max	Paired Wilcoxon test
Initial KS	0	9	8.7	17	V = 6123 p<.001	5	12	10.7	15	V = 72 p<.001	0	9	8.4	17	V = 4746.5 p<.001
Final KS	0	10	9.7	17		4	13.5	12.2	16		0	10	9.4	17	
Difference KS	−7	1	1.1	8		−5	1	1.5	7		−7	1	1.0	8	

Wilcoxon test between difference KS from swing "valid" and swing "invalid": W = 5477, p-value = 1.1e-1 (n.s.).
KS = Knowledge score. KS varies between 0 and 18. Difference KS equals final KS minus initial KS. Difference KS varies between −18 and 18. Min = minimum; Med = median; Max = maximum; p = p-value; n.s. = non-significant.

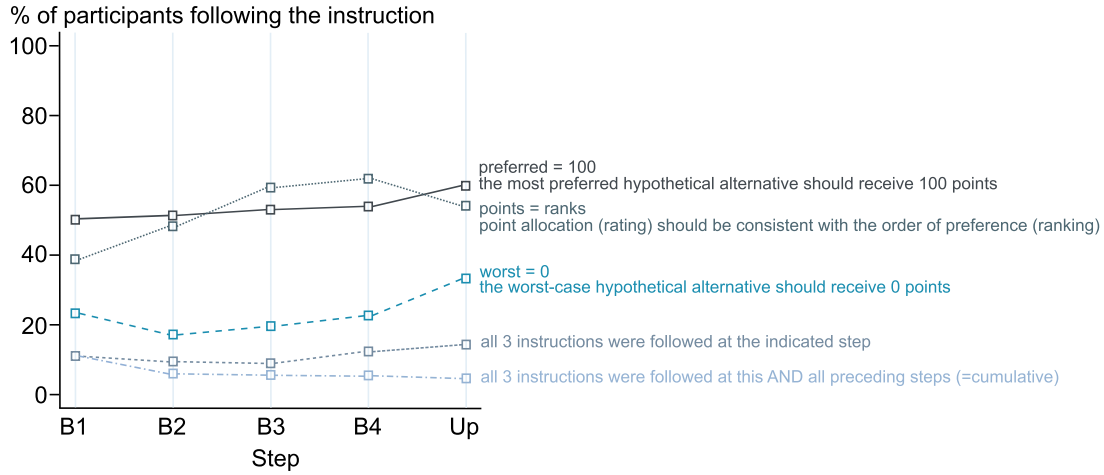

Fig. 4. Process compliance in the online survey to elicit swing weights. Percentage of participants (N = 298) who followed the instructions of the survey. B1-B4: weight elicitation within each branch of the objectives hierarchy (B1 has three objectives, and B2-B4 have two objectives; Fig. 3). Up: elicitation of weights for the upper level-objectives (four upper-level objectives: the most preferred one from each of the four branches). (SI 4b for numerical data).

"preferred = 100". Finally, 15.4% more participants followed the instruction that "the point allocation (rating) should be consistent with the order of preference (ranking) of the hypothetical alternatives" at the last step, compared with the first (Fig. 4, dotted line). The instruction implies that, for instance, an alternative that was ranked in the third position does not receive more points than the alternative ranked in the second position. Overall, considering each instruction separately, more participants were able to follow the instructions at the last step than at the first step. This suggests that participants did learn about the process.

However, the cumulative number of participants following all three instructions together, which was already very low at the start (11.1% in step B1), kept decreasing (Fig. 4, dash-dotted line). At the end, only 14 participants (4.7%) were able to follow all three instructions for all the five steps of the entire swing weight elicitation. In addition, 262 participants (87.9%) gave answers that do not make sense, i.e. they rated at least once at the lower level (steps B1-B4) the dominated worst-case hypothetical alternative (with all objectives on their worst levels) *higher* than one of the other hypothetical alternatives, with one objective at the best possible level. Consequently, in the following, we defined two sub-samples: the "invalid swing" sub-sample composed of these 262 participants, and the "(valid) swing" composed of the remaining 36 participants.

3.4. Preference learning

RQ3: Did participants learn about their preferences during the online survey to elicit swing weights? Overall (N = 298), 151 participants (50.7%) reported that they changed their opinion concerning some of the objectives during the survey, and 207 (69.5%) reported that they strengthened their opinion concerning some objectives. However, many participants reported that they were very uncertain about their answers: on a scale from 0 extremely uncertain to 100 extremely certain, the answers ranged from 1 to 64, with a mean of 22.5, a standard deviation of 20.2, and a median of 28. Thus, although we collected preferences, we must conclude that we cannot be very sure of how well they represent the participants' opinions.

3.5. Is invalid swing a direct rating process?

RQ4: Did the participants who did not comply with the process to elicit swing weights perform a direct rating of objectives? The answer is yes and no: the results for the "invalid swing" participants were always in between those of the participants that correctly carried out the swing survey, and those that filled in the direct rating survey. In other words, we observed a succession in the results from all statistical tests starting with swing, over invalid swing, to direct rating.

Specifically, the spread of weights for individual weight profiles was statistically significantly larger in the swing (mean = 0.14, median = 0.11) than in the invalid swing (mean = 0.10, median = 0.08) (Wilcoxon rank sum test with continuity correction: W = 3363.5, p-value = 5.3e-3). No participant from the group that correctly used the swing weighting gave equal weights (spread of weight = 0), while in the invalid swing group, 2.3% gave equal weights. Note that a post hoc analysis (for Fisher's exact test, carried out with G*Power 3.1.9.4) showed that the power is very low (1 – β = 6.2e-8), suggesting that due to the difference in sample size the results may not be conclusive. Additionally, in the swing group, 52.8% had a spread of weights equal to or higher than 0.11, while in the invalid swing only 37.4% had a spread of weights equal to or higher than 0.11. The results for the direct rating group were even lower in these measures than those of the invalid swing group. In the direct rating survey, the mean spread of weight was 0.06 (median = 0.06), a higher number of participants, i.e. 7.3% gave equal weights, and only 13.4% had a spread of weights equal to or higher than 0.11. The spread of weights for individual weight profiles differed statistically significantly between the invalid swing and direct rating groups (Wilcoxon rank sum test with continuity correction: W = 3406, p-value < 0.001).

We also observed differences in the structure of weights distribution in the relation between the average weight and the rank (Fig. 5). For swing, the relation was concave and relatively steep (collapsed, described by the equation in Table 3). In contrast, the relation was slightly convex and less steep for the invalid swing group (Fig. 5 and Table 3). These curves illustrate that in swing, the difference of weights between two consecutively ranked objectives was higher for the first ranked objectives than for the last ranked objectives. For the invalid swing, the difference of weights between two consecutively ranked objectives was higher for the last ranked objectives than for the first ranked objectives. Note, that similar, but less pronounced, results are observed in the relations between the median weight and the rank (SI 4c). These observations were further supported by the nine independent t-tests comparing the weight distributions for each rank (Table 4). The observations were also supported by the analysis of variance performed with the groups swing and invalid swing as between-participants factor, and the rank as within-participant factor. As expected, there was an effect of rank ($F(8,2368) = 491.6$, $p < .001$), but there was also an effect of the interaction group with the rank ($F(8,2368) = 13.47$, $p < .001$).

The relation between the average weight obtained from the direct rating survey and the rank was convex, with a negligible slope (Fig. 5,

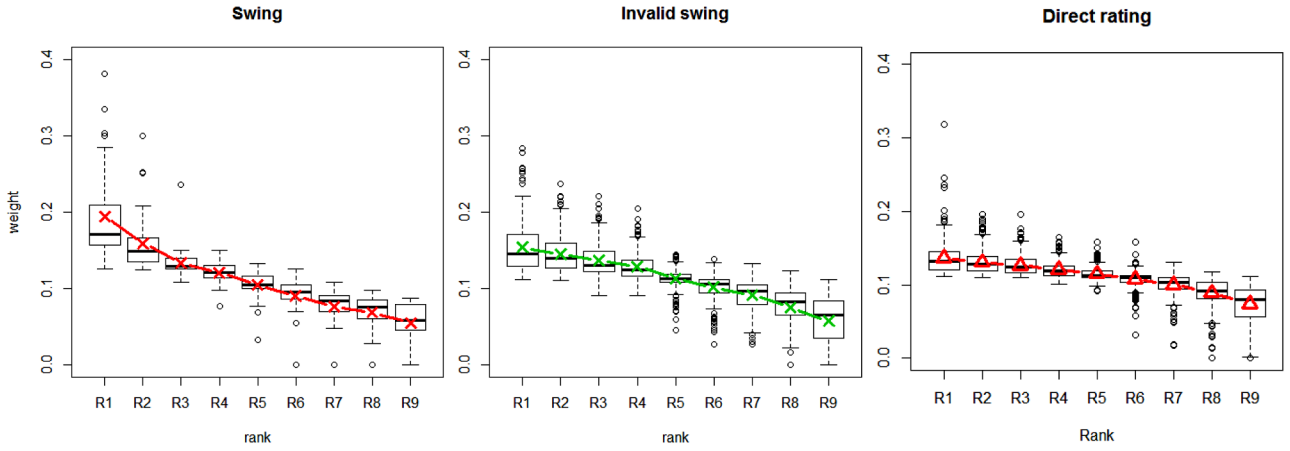

Fig. 5. Spread of weights for the "swing" group (left panel) and the "invalid swing" group (middle), and those who received the direct rating survey (right). The distributions of weights from the highest ranked (R1) to the lowest ranked (R9) objectives are represented with boxplots. The superimposed lines represent the mean following the representation of Doyle, Green and Bottomley [47].

Table 3). The shape of the relation was similar to the one of the invalid swing, however "flatter". The nine independent t-tests comparing the weight distributions for each rank (Table 4) supported this visual observation with statistically significant results. Additionally, the analysis of variance performed with the groups direct rating and invalid swing as between-participants factor, and the rank as within-participant factor again confirmed the statistical difference between the survey types: again, the rank explained some difference in the variance ($F(8,3200) = 947.92$, $p < .001$), and the interaction rank – survey type as well ($F(8,3200) = 18.55$, $p < .001$).

As summary: the weights from the invalid swing group were less spread than those of the group that correctly carried out swing weight elicitation. The weights elicited with the direct rating survey resulted in even less "strong" preferences, i.e. an even lower spread of weights than in the invalid swing group and more equal weights. Moreover, the relation between the average weight and the rank was convex in both the invalid swing group and the group that completed the direct rating survey, consistently with the point allocation curves described by Doyle, Green and Bottomley [47]. In contrast, the swing group had a relatively steep and concave relationship between weights and ranks.

Table 4
Comparing the weights per rank for invalid swing with swing and direct rating. n.s. = non-significant.

	Swing vs. invalid swing			Direct rating vs. invalid swing		
	t	df	p-value	t	df	p-value
Rank1	3.78	37.6	<0.001	−7.22	445	<0.001
Rank2	2.14	38.7	3.87e-2	−8.05	432	<0.001
Rank3	−0.908	46.2	3.69e-1 (n.s.)	−7.00	411	<0.001
Rank4	−2.92	49.9	5.30e-3	−6.07	387	<0.001
Rank5	−2.58	40.2	1.38e-2	2.4	415	1.68e-2
Rank6	−2.54	39.3	1.51e-2	5.41	408	<0.001
Rank7	−3.65	41.5	<0.001	6.47	467	<0.001
Rank8	−1.53	50.0	1.32e-1 (n.s.)	6.58	437	<0.001
Rank9	−0.478	49.7	6.35e-1 (n.s.)	6.77	488	<0.001

4. Discussion

4.1. Answers to our research questions

Our results indicate that participants who answered the online swing surveys somewhat learnt facts about wastewater, but only to a limited extent (RQ1, Table 2). The knowledge score was statistically significantly higher at the end of the survey than at the beginning in both groups. However, an improvement of 1 point (out of 18) hardly indicates substantial factual learning (Table 2). We also found that the participants poorly complied with the instructions of swing weight elicitation process (RQ2, Fig. 4), suggesting low to no process learning. The statistically significant differences in weight patterns between the respondents that complied with the instructions and those that did not, and those that completed the direct rating survey, do indicate that we elicit "something different" with swing (RQ4, Fig. 5, Tables 3 and 4).

However, we still need to understand better in which way a method such as swing supports preference learning in a more suitable way than direct rating [23]. Concerning preference learning (RQ3), half of the swing participants reported that they changed their opinion concerning some objectives during the survey, and 70% that they strengthened some of their pre-existing opinion. However, the participants were highly uncertain about their answers, leading us to conclude that we cannot be very sure of how well these preferences represent the participants' opinions. This uncertainty can also be due to the low understanding of the swing process. The weights expressed by the participants who correctly carried out swing were relatively strongly spread, which indicates more contrasted preferences, especially compared to those participants that carried out the direct rating survey. Those participants that did not comply with the swing instructions for weight elicitation had preferences somewhere in between the patterns observed for the "correct" swing and the direct rating group (RQ4, Figs. 4 and 5, Tables 3 and 4).

A disappointing result is that process compliance was low, and that many participants were not able to follow basic instructions such as giving 0 points to the obviously worst-case hypothetical alternative, where all objectives are on their lowest level (RQ2; Fig. 4). As positive

Table 3
Non-linear curves representing the average weights as a function of the rank. This table accompanies Fig. 5.

	"Swing" group ($N = 36$)			"Invalid swing" group ($N = 262$)			Direct rating ($N = 357$)		
Equation	Weight = 0.217 – 0.0294 rank + 0.001 rank^2			Weight = 0.169 – 0.0060 rank - 0.0006 rank^2			Weight = 0.137 – 1.2e-3 rank - 6.2e-4 rank^2		
	rank	rank^2	adjusted R^2	Rank	rank^2	adjusted R^2	rank	rank^2	adjusted R^2
t(6)	−10.974	5.031	0.9898	−5.630	−5.671	0.99	−1.2	−6.4	0.99
p-value	<0.001	2.38e-3	<0.001	1.34e-3	1.29e-3	<0.001	0.29	<0.001	<0.001

result, the number of participants that complied with each of the individual instructions increased from the first weight elicitation step (B1) to the last step (Up) by 9.7% to 15.5% (depending on the instruction), so some process learning did occur. However, over the entire survey, from weight elicitation in the first branch to weight elicitation at the upper-level of the hierarchy, the cumulative number of participants following all three instructions together, which was already very low at the start (11.1% in step B1) decreased to 4.69% at the last step (Fig. 4). It indicated that some participants, who complied with the instructions at the start, did not comply with the instructions at a later step. Such problems in following the instructions were already reported in the previous online survey to elicit swing weights [30, SI], and we had made large efforts to improve the survey interface based on these experiences. This confirms that unassisted online weight elicitation is not easy [33]. It highlights that there is a real need to program better interfaces to elicit swing weights online, if we (decision analysts) want to meet the praxis demand. Online survey to elicit swing weights – as we did it – was practically feasible, but not well understandable for laypersons. We believe that our study gives interesting insights, allowing us to formulate in which ways we can progress both for improving the design of an online interface to elicit swing weights (Section 4.2) and for future research (Section 4.3).

4.2. Recommendations for future online swing weight elicitation

Our first recommendation for future surveys is to enhance factual learning (RQ1), as participants only increased their knowledge score on average by about one point. For this purpose, tools need to be developed and tested in practical applications. They should be grounded in theory, for instance using practical outputs for instructional design from the research on education and science communication [e.g. 49,50]. Finding the right balance between reducing the text while retaining the necessary information is challenging. The low factual learning observed in our results could also be due to the experiment design: when repeating the knowledge questions at the end, we displayed the participant's initial answers by default. We made this choice because of the length of the survey: our aim was to help the participants. However, this might have created some kind of anchor to the initial answers, particularly if the participants were fatigued at the end of the survey.

Second, the observed low compliance with the instructions for the swing weight elicitation was a striking result (RQ2). Few participants (14, i.e. 4.7%) did follow all of the instructions of the process, but the majority did not (284 respondents, i.e. 95.3%). The previous study applying online swing also reports such issues [30]. There, the authors propose to ignore the inconsistencies in the responses (i.e. rating inconsistent with ranking, rating of the worst-case hypothetical alternative different from 0), if the worst-case hypothetical alternative still receives the lowest rating of all hypothetical alternatives. They report removing 74 answers out of 199 fully completed surveys (37% removed, in the first round of the public survey). In our case, only 36 respondents (12.1%) rated the worst-case hypothetical alternative as the least preferred alternative. This would lead to the deletion of 87.9% of the dataset. Hereafter, we reflect on the interface we proposed.

We aimed to improve the previous interface developed with the standard Qualtrics package [30]. In particular, (1) in the ranking phase, participants could drag and drop the vignettes in their preferred order, instead of writing a number (the rank) in an open box; (2) in the rating phase, participants could see the vignettes ordered according to the rank that they gave, which was not the case in the previous version and was suggested as a cause for the high inconsistency; and (3) in the rating phase, participants used sliders to rate, instead of writing a number (the rate) in an open box. However, these changes were not sufficient. We recommend implementing the following additional features when programming an online survey to elicit swing weights. (1) In the rating phase of the elicitation, not only set the default values of 100 for the most preferred hypothetical alternative, and 0 for the worst-case hypothetical alternative, but fix them (make it impossible for the participants to change the default values of 0 and 100). (2) In the rating phase of the elicitation, do not enable that a lower-ranked hypothetical alternative receives a higher rating than a higher ranked alternative. This means that the slider used for rating (i.e. for giving points between 0 and 100) must automatically stop as soon as it reaches the rating points given to the higher-ranked hypothetical alternative (please note, we already implemented this feature in a prototype for gamifying online weight elicitation to collect reliable preferences [18]). In addition, or alternatively, display a warning message in case of equal ratings. The message would point out that the ratings of the lower-ranked hypothetical alternative are now equal to the alternative on the next-higher rank, and that the rating can be increased only if the rating of the higher-ranked hypothetical alternative is increased. The warning message could also offer the option to repeat the ranking of the hypothetical alternatives in case the participants realise while rating that the ranks do not reflect their preferences after all. (3) Improve the wording of the instructions to make them more prominent, and find the balance between giving sufficient information and examples, and too long and complicated text [44]. This would reduce the risk of tiring the participant, and could make the instructions more understandable. Instructional design should reduce extraneous cognitive load (load created by external interacting elements), and if possible increase the proportion of germane cognitive load (devoted working memory resource) [51]. For instance, the introduction could include a short video explaining the swing weighting process [16]. Alternatively, it should include a "training task" to introduce thinking in terms of preference, followed by a "practice task" to train thinking in term of preferences applied to a familiar topic, as suggested by Anderson and Clemen [46]. (4) Finally, the system should be interactive and ask validation questions, such as consistency check questions, and provide feedback [33,44]. However, such an automatized, personalized procedure requires much more programming than what we have done so far in our surveys. The default standard options in Qualtrics are not sufficient, and presumably, a more flexible program to code surveys is required. For future research, we strongly recommend setting-up and pre-testing different interfaces before programming the actual survey. Additionally, it is highly recommendable to test different design interfaces in controlled experiments.

The two other research questions (RQ3 and RQ4) did not directly lead to recommendations for future online swing weight elicitation. Thus, we discuss them in the next sections.

4.3. Raised research questions

Our results to RQ2 about process compliance not only led to the above list of recommendations, but also raised a series of questions that future research could address. In particular, we observed that the instruction "the point allocation (rating) should be consistent with the order of preference (ranking) of the hypothetical alternatives" displayed the steadiest increase in compliance along the steps of the swing weight elicitation process, with a surprising, substantial drop from the fourth to the fifth step (Fig. 4, dotted line). This sudden drop in the instruction compliance could be an effect of the number of objectives: branches B2 to B4 had only two objectives, while at the highest level of the objectives hierarchy, participants needed to express trade-offs between four objectives simultaneously (Fig. 3). The number of objectives considered simultaneously by the decision-maker was already reported as a factor for inconsistency between preference statements, e.g. by Pöyhönen and Hämäläinen [29]. Future research could focus on rigorously investigating the effect of the number of objectives, in order to make best practice recommendations.

In addition, two partially compensating phenomena seemed to occur. We observed that process compliance increased if we considered each instruction individually (Fig. 4). However, the cumulative process compliance curve kept decreasing. What can this mean? One possible

explanation is that those who complied with the instructions at first might have become tired, and made mistakes towards the end. Future research could monitor those two phenomena at each step (process compliance vs. fatigue), in order to understand what is happening.

More generally, experimental research could focus on the understandability and the concrete design of the process. For instance, one could test how well participants understand the vignettes representing the hypothetical alternatives, including the dominated worst-case alternative, where all objectives are on their worst level, e.g. versus more simple graphs. That would inform us of whether participants understand the proposed graphical representation, and what a dominated alternative is. In addition to experimentally testing the vignettes themselves, research could focus on the manipulation of those vignettes. Is drag and drop of vignettes in a list the best design for ranking, in comparison to, for instance, drag and drop from a first list to a second list, or assigning a rank number in a text box, or assigning a rank number in a matrix, etc.? Similarly, rating could be done by other means than sliders. To address these points, the literature on survey design could be useful [e.g. 52,53]. Moreover, future work could focus on how participants perceive the survey, in relation to their personal characteristics (e.g. age, level of systems intelligence [54], etc.).

We had used process compliance as a proxy for process learning. Specifically testing for process learning would have required developing additional targeted test and self-reported questions in the feedback part of our already long survey. We encourage further experimental research, as opposed to our real-world application, to investigate along those lines, building up on the literature on measuring constructed preferences [44,46]. Practically, studying the process of weight elicitation itself could be achieved by, for instance, asking participants to comment aloud on what they are doing or thinking while completing the survey, i.e. using think-aloud protocols [55]. A simpler approach would be to develop a post-questionnaire aimed at unravelling the participants' behaviour during the swing survey. These studies would enable us to assess if the preference reversal observed between the ranking and the rating phase is a real phenomenon, as known from the behavioural economics literature [e.g. 56], or if it is solely a consequence of a lack of attention or tiredness of the participants.

Addressing those questions would also help to better assess preference construction [57], going beyond the measure of constructed preference [44]. In particular, understanding why participants did not comply with the instructions would inform us about how to deal with the collected data to calculate the weights. Did they lower the rating of the most preferred alternative by mistake? Or did they mean to do so because they were considering a global scale [58]? Would it make sense in that case, to rescale the ratings, e.g. stretching the whole set of ratings so that the highest rating equals 100? Note, that this would systematically contribute to an increased spread of weights. We observed a statistically significantly stronger spread of weights amongst those participants that used swing than those of the "invalid" swing and the direct rating group (RQ4). More in-depth studies concerning the spread of weights in general might be informative, since it is known that the spread of weights can vary amongst elicitation procedures, and that e.g. AHP can cause a larger spread [29], but possibly also the smart/swing variant, compared to swing [30]. It would also be interesting to estimate how much of the spread of weights is produced by the hierarchical weighting procedure.

Thus, we need theoretical work that investigates how to measure the construction of preference. Developing standardized protocols to measure the process of preference construction would help decision science to progress. It would help disentangling whether the observed spread of weights is actually due to methodological choices (e.g. elicitation method, hierarchical elicitation technique, calculation to transform the ratings into weights, etc.), or actually represents the persons' preferences.

Moreover, the theoretical work on preference construction could also investigate whether and how tasks perceived as more demanding would enable respondents to construct and stabilize preferences more rapidly than tasks perceived as less demanding (e.g. online swing weight elicitation vs. direct-rating), as suggested by Hoeffler and Ariely [59], but challenged by our results. That would be a first step in attempting to explain RQ4, where our results suggest that participants who did not comply with the swing instructions seemed to have done something more similar to direct rating.

All the above suggestions indicate that there are many opportunities for exciting – and relevant – research. We think that both controlled field studies and well-design experimental lab research can be especially interesting for the young field of Behavioural Operational Research (BOR) [60,61].

4.4. Limitations of the present study

The present study assessed an online survey to elicit swing weights, developed on a basic commonly used survey platform. It focused on the swing method (commonly used in MAVT/MAUT), and was constrained by the supporting platform. In addition, because it was a real-world application with affected citizens, the experimental design was limited. Thus, criticism is legitimate. Most likely, the weakest point is the lack of proper measures to assess preference learning (RQ3), and process learning (RQ2). One could also point out that our measure of factual learning (RQ1) as a pre- and post-knowledge test would rather be an assessment of the short-term memory. Finally, the recruitment of the participants through a panel company may also induce some biases. However, our aim matched those limitations: we aimed to test the feasibility, i.e. the practicality and understandability, of the proposed online interface to elicit swing weights in a real-world application. We hope that reporting on the weaknesses and the negative results contributes to (1) overcome publication and confirmation biases [62,63], which can lead to overrate treatments and distort results of meta-analyses, (2) establish good practice guidelines (or at least many recommendations), and (3) emphasize research opportunities.

4.5. Reflecting on using online swing to increase citizen participation

As presented in the introduction, praxis is demanding increased engagement of citizens for dealing with complex environmental issues. In order to do this, we need some way of understanding the public's preferences. Two aspects are specific to using online swing weight elicitation, as a means to increase citizen participation [64]. First, citizens are confronted with Value-Focused Thinking [25], i.e. they need to think in terms of trade-offs between objectives. This differs from most real-world decision-making, where citizens express their preferences for alternatives, for instance by voting for or against a solution in a public referendum (consultative mode). Second, when using online swing weight elicitation, there is a two-directional exchange between the citizens and the initiator of the survey, i.e. the initiator communicates facts concerning the environmental decision problem, and the citizens share their preferences. However, "it is very possible that the 'information' displayed on a DSS [Decision Support System] does not truly inform the user", as suggested by French [9]. French goes on questioning whether the computer actually supports the participants cognitively in their task, and whether completing this task supports the overall decision-making process. We thus need to reflect (and test) good ways to inform users.

We believe that a careful, well-designed use of MCDA methods, including swing weight elicitation, can address such problems regarding broad public participation via online DSS. Through the use of online swing weight elicitation, a structured decision-making process could not only emphasize the depth of the process (information on the problem, knowledge, and values are thoroughly considered) [64], but also its breadth (many participants can take part in the process). Online elicitation processes allow a large number of people to be involved.

Table 5
Comparing the advantages and challenges of online vs. facilitated citizen participation (face-to-face interviews or group workshops). Summary of the arguments from the introduction and discussion, where relevant references can be found.

	Advantages	Challenges
Online participation	• Is requested from praxis/ citizens to allow broad stakeholder and public participation • Potentially allows involving all affected individuals, including laypersons • Complies with "participatory democracy" • Is easily accessible; can be carried out anyplace, anytime • Strongly broadens range of opinions, thus making potential conflict lines more obvious • If set up by an experienced decision analyst, can be carried out by less experienced analysts • Can be easily adapted to different settings • Can be easily replicated in different cases, opening up interesting research opportunities	• Is limited to people that have the technical and cognitive means to use online device (computer, tablet, etc.) • Is potentially demotivating (high cognitive demand, relatively long surveys, no external motivation from facilitator) • Requires well-designed interfaces to enhance process learning without facilitation • Increases the risk of people using heuristics and running into biases, which leads to the risk of producing preference data of disputable quality • Needs that we (researchers) find ways to control the occurrence of heuristics and biases
Face-to-face participation	• Facilitates a shared understanding of the decision at stake • Favours finding consensual solution • Complies with "deliberative democracy" • Thanks to the direct interaction with facilitator, guides participants to think in a structured, consistent way • Thanks to the direct interaction, especially with an experienced facilitator, allows for extensive consistency checks and noticing possible biases • Thanks to the direct interaction with facilitator, may be perceived as less repetitive and boring than working on one's own	• Allows involving only few selected stakeholders (e.g., 15 to 20 in a workshop or in face-to-face interviews) • Is time-consuming for (busy) participants and facilitator • Is space constrained (e.g., room) • Requires finding a time slot in the agenda that suits all involved • Requires skilled and well-trained facilitators to enhance process learning • Potentially represents a too narrow range of opinions, if participants are not very carefully selected • Can potentially lead to losing individual participants' opinions in group workshop • Can potentially lead to the group thinking bias in group workshop

However, many meta-choices are required to develop the functional interface [65]: these are still challenging researchers and practitioners. Finally, Table 5 summarizes possible advantages and disadvantages of an online process, versus face-to-face involvement (in individual interviews or group workshops).

5. Conclusion

Our present study aimed to assess the feasibility, i.e. the practicality and understandability, of eliciting swing weights through an online interface, in a real-world application with affected citizens. We improved a previously developed interface. We assessed the feasibility, based on three aspects of learning in the context of MCDA: individual factual learning, process compliance (following the instructions), and preference learning. In total, 655 inhabitants from the Paris region answered our online surveys to elicit swing weights, or to directly rate objectives for comparison purposes. While some factual learning occurred, it was limited. The proportion of participants not complying with the swing process was very high, as observed in a previous real-world attempt for online weight elicitation [30]. Participants answering the online swing weight elicitation survey who complied with the instructions had more pronounced preferences (the weights were more spread), compared to those that did not comply with the swing instructions, and especially those that answered the direct rating survey. A majority of the participants reported that they had changed or strengthened their existing preferences during the survey to elicit swing weights, which indicates that some preference learning occurred. However, because they also stated that they were uncertain about their answers, we are unsure of how well these elicited preferences represent the participants' opinions.

Encountering the many issues in this survey enabled us to formulate a series of practical recommendations for designing interfaces for future online swing weight elicitation, which aim at enhancing factual learning and process compliance. The encountered problems also allowed us to highlight a number of research questions, whose answers will contribute to improving the design and implementation of online swing weight elicitation. Additionally, we believe that some of the proposed research questions are of general importance for MCDA researchers, and especially for the new stream of Behavioural Operational Research (BOR). Our propositions (extensively developed in Section 4.3) include: (1) exploring the number of objectives that can be simultaneously assessed in elicitation, (2) understanding the cognitive processes and the effects of different weighting procedures and methods, (3) exploring how fatigue of respondents might reduce process compliance, (4) researching visual representations and interactive designs, and more. Controlled experiments, with the methods or interfaces as varying factors, and/or think-aloud protocols should help answering those questions, as well as further real-world applications. Finally, and probably as largest challenge, we deem it as crucial that we explore better measures for understanding process learning and preference construction. We hope that reporting our experience will help building a community of practice and supports developing a practical guide for online swing weight elicitation. This is in the interest of practitioners and researchers alike, who are concerned with engaging a larger number of citizens in complex environmental and public policy decisions.

CRediT authorship contribution statement

Alice H. Aubert: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing - original draft, Visualization, Supervision, Funding acquisition. **Fabien Esculier:** Resources, Writing - review & editing, Project administration, Funding acquisition. **Judit Lienert:** Resources, Writing - review & editing, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work is the result of a collaboration between Eawag and the OCAPI program lead by ENPC (www.leesu.fr/ocapi). The authors

warmly thank the participants in the OCAPI research program. The authors express their special thanks to participants of the stakeholder workshop, to Tove Larsen and Kai Udert at Eawag for providing engineering expertise, and to the proof-readers of the survey and its translation to French. We are very grateful to the editor and three anonymous reviewers for their support and excellent guidance in improving this manuscript.

Funding

This work was supported by Eawag, the Swiss Federal Institute of Aquatic Science and Technology [Grant No. 5221.00492.009.08 DF 15, 2015], the Syndicat Interdépartemental d'Assainissement de l'Agglomération Parisienne, the Agence de l'Eau Seine Normandie, the École des Ponts ParisTech and the Ministère de la Transition Écologique et Solidaire [OCAPI program – phase I (www.leesu.fr/ocapi)].

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.orp.2020.100156.

References

[1] Papadopoulos Y, Warin P. Are innovative, participatory and deliberative procedures in policy making democratic and effective? Eur J Polit Res 2007;46:445–72. https://doi.org/10.1111/j.1475-6765.2007.00696.x.
[2] Irvin RA, Stansbury J. Citizen participation in decision making: is it worth the effort. Public Adm Rev 2004;64:55–65. https://doi.org/10.1111/j.1540-6210.2004.00346.x.
[3] Euler J, Heldt S. From information to participation and self-organization: visions for European river basin management. Sci Total Environ 2018;621:905–14. https://doi.org/10.1016/j.scitotenv.2017.11.072.
[4] Luhrs N, Jager NW, Challies E, Newig J. How Participatory should environmental governance be? testing the applicability of the Vroom-Yetton-Jago model in public environmental decision-making. Environ Manage 2018;61:249–62. https://doi.org/10.1007/s00267-017-0984-3.
[5] Reed MS, Vella S, Challies E, de Vente J, Frewer L, Hohenwallner-Ries D, Huber T, Neumann RK, Oughton EA, del Ceno JS, et al. A theory of participation: what makes stakeholder and public engagement in environmental management work. Restorat Ecol 2018;26:S7–17. https://doi.org/10.1111/rec.12541.
[6] Gray S, Voinov A, Paolisso M, Jordan R, BenDor T, Bommel P, Glynn P, Hedelin B, Hubacek K, Introne J, et al. Purpose, processes, partnerships, and products: four Ps to advance participatory socio-environmental modeling. Ecol Appl 2018;28:46–61.
[7] Luyet V, Schlaepfer R, Parlange MB, Buttler A. A framework to implement Stakeholder participation in environmental projects. J Environ Manage 2012;111:213–9. https://doi.org/10.1016/j.jenvman.2012.06.026.
[8] Brown G, Chin SYW. Assessing the effectiveness of public participation in neighbourhood planning. Plan Pract Res 2013;28:563–88. https://doi.org/10.1080/02697459.2013.820037.
[9] French S. Web-enabled strategic GDSS, e-democracy and Arrow's theorem: a Bayesian perspective. Decis Support Syst 2007;43:1476–84. https://doi.org/10.1016/j.dss.2006.06.003. doi:https://doi.org/.
[10] Munda G. Social multi-criteria evaluation for a sustainable economy. Heidelberg: Springer-verlag Berlin; 2008. https://doi.org/10.1007/978-3-540-73703-2pp. XVII, 210.
[11] Baker S, Chapin Iii FS. Going beyond "it depends": the role of context in shaping participation in natural resource management. Ecol Soc 2018;23. https://doi.org/10.5751/ES-09868-230120.
[12] Garmendia E, Stagl S. Public participation for sustainability and social learning: concepts and lessons from three case studies in Europe. Ecol Econ 2010;69:1712–22. https://doi.org/10.1016/j.ecolecon.2010.03.027.
[13] Redpath SM, Keane A, Andrén H, Baynham-Herd Z, Bunnefeld N, Duthie AB, Frank J, Garcia CA, Månsson J, Nilsson L, et al. Games as Tools to Address Conservation Conflicts. Trends Ecol Evol (Amst) 2018;33:415–26. https://doi.org/10.1016/j.tree.2018.03.005.
[14] Harris-Lovett SR, Binz C, Sedlak DL, Kiparsky M, Truffer B. Beyond user acceptance: a legitimacy framework for potable water reuse in california. Environ. Sci. Technol. 2015;49:7552–61. https://doi.org/10.1021/acs.est.5b00504.
[15] Liu S, Maclean K, Robinson C. A cost-effective framework to prioritise stakeholder participation options. EURO J Decis Process 2019;7(4):221–41. https://doi.org/10.1007/s40070-019-00103-7. in press.
[16] Mustajoki J, Hämäläinen RP, Marttunen M. Participatory multicriteria decision analysis with Web-HIPRE: a case of lake regulation policy. Environ Modell Software 2004;19:537–47. https://doi.org/10.1016/j.envsoft.2003.07.002.
[17] Gregory R, Satterfield T, Hasell A. Using decision pathway surveys to inform climate engineering policy choices. Proceed Natl Acad Sci 2016;113:560–5. https://doi.org/10.1073/pnas.1508896113.
[18] Aubert AH, Lienert J. Gamified online survey to elicit citizens' preferences and enhance learning for environmental decisions. Environ Modell Software 2019;111:1–12. https://doi.org/10.1016/j.envsoft.2018.09.013.
[19] Marttunen M, Belton V, Lienert J. Are objectives hierarchy related biases observed in practice? A meta-analysis of environmental and energy applications of Multi-Criteria Decision Analysis. Eur J Oper Res 2018;265:178–94. https://doi.org/10.1016/j.ejor.2017.02.038.
[20] Huang IB, Keisler J, Linkov I. Multi-criteria decision analysis in environmental sciences: ten years of applications and trends. Sci Total Environ 2011;409:3578–94. https://doi.org/10.1016/j.scitotenv.2011.06.022.
[21] Keeney RL. Value-focused thinking: identifying decision opportunities and creating alternatives. Eur J Oper Res 1996;92:537–49. https://doi.org/10.1016/0377-2217(96)00004-5.
[22] Gregory R, Failing L, Harstone M, Long G, McDaniels TL, Ohlson D. Structured decision making: a practical guide to environmental management choices. West Sussex, UK: Wiley-Blackwell; 2012. p. 299.
[23] Eisenführ F, Weber M, Langer T. Rational decision making. Germany: Springer Berlin Heidelberg; 2010. Berlin Heidelberg, GermanyXIV, 447.
[24] Langhans SD, Reichert P, Schuwirth N. The method matters: a guide for indicator aggregation in ecological assessments. Ecol Indic 2014;45:494–507. https://doi.org/10.1016/j.ecolind.2014.05.014.
[25] Keeney RL, Raiffa H. Decisions with multiple objectives: preferences and value tradeoffs. New York: Wiley; 1976. p. 569.
[26] Riabacke M, Danielson M, Ekenberg L. State-of-the-art prescriptive criteria weight elicitation. Adv Decis Sci 2012;2012:1–24. https://doi.org/10.1155/2012/276584.
[27] Phillips LD. Decision Conferencing. In: von Winterfeldt D, Miles Jr RF, Edwards W, editors. Advances in decision analysis: from foundations to applications Cambridge: Cambridge University Press; 2007. p. 375–99. https://doi.org/10.1017/CBO9780511611308.020.
[28] Lourenco RP, Costa JP. Incorporating citizens' views in local policy decision making processes. Decis Support Syst 2007;43:1499–511. https://doi.org/10.1016/j.dss.2006.06.004.
[29] Pöyhönen M, Hämäläinen RP. On the convergence of multiattribute weighting methods. Eur J Oper Res 2001;129:569–85. https://doi.org/10.1016/S0377-2217(99)00467-1.
[30] Lienert J, Duygan M, Zheng J. Preference stability over time with multiple elicitation methods to support wastewater infrastructure decision-making. Eur J Oper Res 2016;253:746–60. https://doi.org/10.1016/j.ejor.2016.03.010.
[31] Montibeller G, von Winterfeldt D. Cognitive and motivational biases in decision and risk analysis. Risk Anal 2015;35:1230–51. https://doi.org/10.1111/risa.12360.
[32] van Til J, Groothuis-Oudshoorn C, Lieferink M, Dolan J, Goetghebeur M. Does technique matter; a pilot study exploring weighting techniques for a multi-criteria decision support framework. Cost Eff Resour Allocation 2014;12:22. https://doi.org/10.1186/1478-7547-12-22.
[33] Marttunen M, Hämäläinen RP. The decision analysis interview approach in the collaborative management of a large regulated water course. Environ Manage 2008;42:1026–42. https://doi.org/10.1007/s00267-008-9200-9.
[34] Zheng J, Lienert J. Stakeholder interviews with two MAVT preference elicitation philosophies in a Swiss water infrastructure decision: aggregation using SWING-weighting and disaggregation using UTAGMS. Eur J Oper Res 2018;267:273–87. https://doi.org/10.1016/j.ejor.2017.11.018.
[35] Hämäläinen RP. Behavioural issues in environmental modelling – The missing perspective. Environ Modell Software 2015;73:244–53. https://doi.org/10.1016/j.envsoft.2015.08.019.
[36] Voinov A, Kolagani N, McCall MK, Glynn PD, Kragt ME, Ostermann FO, et al. Modelling with stakeholders – Next generation. Environ Modell Software 2016;77:196–220. https://doi.org/10.1016/j.envsoft.2015.11.016.
[37] Hämäläinen R, Kettunen E, Marttunen M, Ehtamo H. Evaluating a framework for multi-stakeholder decision support in water resources management. Group Decis Negotiat 2001;10:331–53. https://doi.org/10.1023/A:1011207207809.
[38] Belton V, Elder MD. Decision support systems: learning from visual interactive modelling. Decis Support Syst 1994;12:355–64. https://doi.org/10.1016/0167-9236(94)90052-3.
[39] Belton V, Branke J, Eskelinen P, Greco S, Molina J, Ruiz F, Słowiński R. Interactive Multiobjective Optimization from a Learning Perspective. In: Branke J, Deb K, Miettinen K, Słowiński R, editors. Multiobjective optimization: interactive and evolutionary approaches Berlin, Heidelberg: Springer Berlin Heidelberg; 2008. p. 405–33. https://doi.org/10.1007/978-3-540-88908-3_15.
[40] Monks T, Robinson S, Kotiadis K. Learning from discrete-event simulation: exploring the high involvement hypothesis. Eur J Oper Res 2014;235:195–205. https://doi.org/10.1016/j.ejor.2013.10.003.
[41] Thompson JP, Howick S, Belton V. Critical learning incidents in system dynamics modelling engagements. Eur J Oper Res 2016;249:945–58. https://doi.org/10.1016/j.ejor.2015.09.048.
[42] Lichtenstein S, Slovic P. The Construction Of Preference: an Overview. In: Lichtenstein S, Slovic P, editors. The construction of preference. New York: Cambridge University Press; 2006. p. 1–40.
[43] Mezirow J. Learning as transformation: critical perspectives on a theory in progress. San Francisco: Jossey Bass; 2000.
[44] Payne JW, Bettman JR, Schkade DA. Measuring Constructed Preferences: towards a Building Code. In: Lichtenstein S, Slovic P, editors. The construction of preference. New York: Cambridge University Press; 2006. p. 629–52.
[45] Esculier, F.; Aubert, A.H.; Lienert, J.; Larsen, T.A.Selection and assessment of criteria to evaluate scenarios of urine and faeces urban management. in prep.
[46] Anderson RM, Clemen R. Toward an improved methodology to construct and reconcile decision analytic preference judgments. Decis Anal 2013;10:121–34.

https://doi.org/10.1287/deca.2013.0268.
[47] Doyle JR, Green RH, Bottomley PA. Judging Relative Importance: direct Rating and Point Allocation Are Not Equivalent. Organ Behav Hum Decis Process 1997;70:65–72. https://doi.org/10.1006/obhd.1997.2694.
[48] R Development Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2017. URL http://www.R-project.org/.
[49] Littledyke M. Science education for environmental awareness: approaches to integrating cognitive and affective domains. Environ Educ Res 2008;14:1–17. https://doi.org/10.1080/13504620701843301.
[50] Plass JL, Homer BD, Kinzer CK. Foundations of game-based learning. Educ Psychol 2015;50:258–83. https://doi.org/10.1080/00461520.2015.1122533.
[51] Sweller J. Cognitive load theory. In: Seel NM, editor. Encyclopedia of the sciences of learning. Dordrecht Heidelberg London: Springer: New York; 2012. 10.1007/ 978-1-4419-1428-6pp. 2182-2184.
[52] Dillman DA, Smyth JD, Christian LM. Internet, mail, and mixed-mode surveys: the tailored design method. 3rd ed. Hoboken, New Jersey, USA & Canada: John Wiley & Sons, Inc.; 2009. p. 500.
[53] Funke F. A Web Experiment Showing Negative Effects of Slider Scales Compared to Visual Analogue Scales and Radio Button Scales. Soc Sci Comput Rev 2016;34:244–54. https://doi.org/10.1177/0894439315575477.
[54] Hämäläinen RP, Saarinen E. Systems intelligence - Discovering a hidden competence in human action and organizational life. Finland: Helsinki University of Technology - Systems Analysis Laboratory; 2004.
[55] Jääskeläinen R. Think-aloud protocol. In: Gambier Y, van Doorslaer L, editors. Handbook of translation studies, 1. Amsterdam/ Philadelphia: John Benjamins Publishing Company; 2010. p. 371–3.
[56] Alós-Ferrer C, Granić Đ-G, Kern J, Wagner AK. Preference reversals: time and again. J Risk Uncertain 2016;52:65–97. https://doi.org/10.1007/s11166-016-9233-z.
[57] Slovic P. The construction of preference. American Psychologist 1995;50:364–71. https://doi.org/10.1037/0003-066X.50.5.364.
[58] Monat JP. The benefits of global scaling in multi-criteria decision analysis. Judgm Decis Mak 2009;4:492–508.
[59] Hoeffler S, Ariely D. Constructing stable preferences: a look into dimensions of experience and their impact on preference stability. Journal of Consumer Psychology 1999;8:113–39. https://doi.org/10.1207/s15327663jcp0802_01.
[60] Franco LA, Hämäläinen RP. Behavioural operational research: returning to the roots of the OR profession. Eur J Oper Res 2016;249:791–5. https://doi.org/10.1016/j.ejor.2015.10.034.
[61] Hämäläinen RP, Luoma J, Saarinen E. On the importance of behavioral operational research: the case of understanding and communicating about dynamic systems. Eur J Oper Res 2013;228:623–34. https://doi.org/10.1016/j.ejor.2013.02.001.
[62] Sridharan L, Greenland P. Editorial Policies and Publication Bias: the Importance of Negative Studies. Arch. Intern. Med. 2009;169:1022–3. https://doi.org/10.1001/archinternmed.2009.100.
[63] Moahoney MJ. Publication prejudices: an experimental study of confirmatory bias in the peer review system. Cognit Ther Res 1977;1:161–75.
[64] Gregory R, Failing L, Hardstone M, Long G, McDaniels T, Ohlson D. Reality Check: implementation (Chapter 11). Structured decision making: a practical guide to environmental managament choices. Wiley-Blackwell; 2012. p. 263–88.
[65] Ferretti V, Montibeller G. Key challenges and meta-choices in designing and applying multi-criteria spatial decision support systems. Decis Support Syst 2016;84:41–52. https://doi.org/10.1016/j.dss.2016.01.005.