Stakeholder interviews with two MAVT preference elicitation philosophies in a Swiss water infrastructure decision: aggregation using SWING-weighting and disaggregation using UTA^{GMS}

Jun Zheng[a,*], Judit Lienert[a]

[a] *Swiss Federal Institute of Aquatic Science and Technology, Eawag, P.O. Box 611, 8600 Duebendorf, Switzerland*

[*] Corresponding author

Email addresses: `zhengjun516@hotmail.com` (Jun Zheng), `judit.lienert@eawag.ch` (Judit Lienert)

Abstract We used two types of preference elicitation methods based on multi-attribute value theory (MAVT) for a wastewater infrastructure decision in Switzerland. We aimed to register the implementation impacts of two preference elicitation philosophies (aggregation, disaggregation) in a large, real-world case and give guidance on these elicitation approaches for practitioners. We conducted two series of face-to-face interviews with the same ten. The first interview set used direct aggregation preference elicitation methods, which decomposed an additive value model into the elicitation of weights (SMART/SWING-variant) and marginal value functions (bi-section method). In the second interview series, indirect disaggregation was used, based on UTA$^{\text{GMS}}$. The weights and marginal value functions for 19 objectives were later simultaneously inferred with linear programming from pairwise comparisons of hypothetical alternatives. One aim was to design the UTA$^{\text{GMS}}$ comparisons for many objectives. Further, we aimed to identify differences and commonalities of the two methods concerning the elicited preferences, the MAVT evaluation results of six real-world wastewater infrastructure alternatives, and the stakeholders' and analysts' feedbacks. Similar best alternatives indicate convergence of the two elicitation methods. This demonstrates the applicability of the UTA$^{\text{GMS}}$ elicitation procedure to a very complex decision problem. However, the two elicitation methods were perceived differently by the respondents and required different effort from the analysts. For individual stakeholders, preferences were sometimes rather different between the interviews, which could be largely explained by the constructive nature of preference formation. This indicates the importance of supporting stakeholder learning in the application of MCDA.

Keywords: Multiple criteria analysis, Behavioral OR, Preference elicitation, OR in environment and climate change, Stakeholder interview.

1. Introduction

Multi-Criteria decision analysis (MCDA) aims at supporting decision makers to evaluate alternatives on several conflicting criteria (Figueira et al., 2016). Incorporating the decision makers' values is crucial, and the preference elicitation process should meaningfully represent the stakeholders' decision preferences by specific MCDA methods and their associated parameters. MCDA can be considered an "umbrella term to describe a collection of formal approaches which seek to take explicit account of multiple criteria in helping individuals or groups explore decisions that matter" (Belton and Stewart, 2002). These formal MCDA approaches include Multi-attribute value and utility theory (MAVT/ MAUT) (Keeney and Raiffa, 1976), the Analytic Hierarchy Process (AHP) (Saaty, 1980), the Dominance-based rough set approach (Greco et al., 2001), the family of outranking methods (Roy, 1996), and others.

Every approach has weaknesses and strengths (Cinelli et al., 2014). In practice, presumably the most popular MCDA method is AHP (see e.g. review by Marttunen et al., 2017), but AHP has been repeatedly heavily criticized (e.g. Macharis et al., 2004; Smith and von Winterfeldt, 2004). AHP is attractive, because the required pairwise comparisons seem to mimic the way people intuitively make decisions. Outranking methods are also very widely used in practice and certainly provide a valid basis for MCDA. These methods also use pairwise comparisons and outranking relations (for a comprehensive overview see Figueira et al., 2016). There is user-friendly software available, continuously developed by promoters of ELECTRE (e.g. Figueira et al., 2013; Roy, 1996; Roy and Bouyssou, 1993) or PROMETHEE-GAIA (e.g. Behzadian et al., 2010; Brans et al., 1986). Our reasons for choosing the widely applied MAVT have been specified in detail by Reichert et al. (2015) and Schuwirth et al. (2012): In environmental decisions, the method and decision process need to be justifiable to the public, and should thus be as transparent as possible. The method should allow easy integration and quantification of best-available scientific knowledge as well as of the large uncertainty that stems from the prediction of decision outcomes in complex ecological and engineered systems. The mathematical formalism of MAVT satisfies these key conceptual requirements by being based on few, but solid rationality axioms, by allowing easy modeling of uncertainty and updating with new information, and by giving large mathematical freedom in describing stakeholder preferences.

Traditionally in MAVT, the preference parameters required for the mathematical models are elicited separately. These include marginal value functions, importance weights, and additional parameters in the (rare) cases of non-additive aggregation (Langhans and Lienert, 2016; Langhans et al., 2014). Many methods have been proposed for separately eliciting marginal value functions and weights (overviews see e.g. Eisenführ et al., 2010; Keeney and Raiffa, 1976; Pöyhönen and Hämäläinen, 2001). This separate elicitation of preference parameters is referred to as direct aggregation or decomposed methods (Beinat, 1997; Doumpos and Zopounidis, 2011). In the application of MAVT in real-world decisions, the weights and marginal value functions are usually separately elicited following this direct aggregation philosophy. Environmental applications range from wastewater treatment (Lienert et al., 2011), over forest management (Mustajoki et al., 2011), to the reuse of historical heritage (Ferretti et al., 2014), and they involve interaction with real stakeholders or decision makers (reviews e.g. Gregory et al., 2012; Hajkowicz and Collins, 2007; Huang et al., 2011; Mendoza and Martins, 2006).

Indirect disaggregation methods have been proposed to infer MAVT models from a set of decision examples on some reference alternatives (Jacquet-Lagreze and Siskos, 2001). The UTA method was the first proposed to infer an additive value function from a ranking of reference alternatives (Jacquet-Lagreze and Siskos, 1982). The piecewise linear value functions are estimated with ordinal regression using linear programming. Due to insufficient preference information, a unique model can hardly be determined. Several UTA variants have then been presented, such as UTASTAR and ACUTA, using some criterion to select one single model (Beuthe and Scannella, 2001; Bous et al., 2010; Siskos et al., 2016). Robust ordinal regression (ROR) is another way of dealing with non-uniqueness, taking into account all possible value functions compatible with the preferences without selecting one particular model. $\mathrm{UTA}^{\mathrm{GMS}}$ or GRIP are two possible methods (Figueira et al., 2009; Greco et al., 2008). Research on the disaggregation methods is currently very active (e.g. Corrente et al., 2012; Greco et al., 2014; Kadziński and Tervonen, 2013), applications are reported in many diverse fields. These range from financial management (e.g. Doumpos et al., 2001; Zopounidis, 2001; Zopounidis et al., 2007), over healthcare (Doumpos et al., 2016), to brand image (Ghaderi et al., 2015). Some studies involve direct interaction with stakeholders to elicit their preferences, e.g. a job evaluation process carried out in a large Greek organization (Spyridakos et al., 2001). Environmental applications typically aim to develop useful models for decision-making, often using real-world data. Examples concern assessing the environmental impact of European cities (Kadziński et al., 2016), or influences on the energy effectiveness of countries (Diakoulaki et al., 1999). However, most environmental UTA studies that we are aware of use historical decisions, but rarely involved direct and active interaction with stakeholders to elicit their preferences. A notable exception is a recent assessment to increase sustainability in the production of silver nanoparticles, where preference information was collected from two chemists (Kadziński et al., 2016).

To date it is still unclear, which approach is actually a good choice to support real-world decisions. We argue that it is highly desirable that interactions with stakeholders applying different elicitation methods or MCDA paradigms produce roughly the same results (e.g. Anderson and Clemen, 2013). In other words: in real-world decisions we would strongly hope that the decision analyst can give the same recommendation concerning the best alternatives, regardless of the chosen methods. While intuitively appealing, this is by no means certain. There is some research available from behavioral decision analysis that has directly compared different elicitation approaches; especially weight elicitation methods (e.g. Belton, 1986; Borcherding et al., 1991; Lienert et al., 2016; Mustajoki et al., 2005; Pöyhönen and Hämäläinen, 2001; Weber and Borcherding, 1993). These comparative studies of elicitation methods are mainly experimental, and to our knowledge so far exclude disaggregation methods of the UTA family (Weber and Borcherding, 1993). Researchers of the disaggregation methods claim that these require less cognitive effort from the decision maker (e.g. Branke et al., 2017; Greco et al., 2008; Kadziński and Tervonen, 2013; van Valkenhoef and Tervonen, 2016), and follow people's natural way of reasoning (e.g. Kadziński et al., 2016). But there is limited evidence to support this claim. On the other hand, for complex decisions that involve many objectives (or many alternatives if these are directly compared), a large number of questions is needed for pairwise comparison approaches. This can become tiring for decision makers, and thus also cognitively demanding (e.g. Macharis et al., 2004; van Valkenhoef and Tervonen, 2016).

More generally, we are still far from understanding in which way different methods, different framings, and different representations of the decision problem, etc. actually influence preferences and thus the decision outcome. A large body of research of behavioral psychologists strongly indicates that preferences are constructed during and influenced by the decision-making process (reviews see e.g. Lichtenstein and Slovic, 2006; Payne et al., 1992; Slovic, 1995). Recently, various authors from Behavioral Operational Research (BOR: Franco and Hämäläinen, 2016; Hämäläinen et al., 2013) and from environmental modelling (Hämäläinen, 2015; Voinov et al., 2016) strongly advocate that we do not only need exciting and innovative theory-framed material. Additionally, research should also focus on understanding in which way relevant and robust outcomes are produced in real-world interventions. They also advocate that current best practice to increase the robustness of results from decision-making interventions is to use multiple case studies and framings, and multiple elicitation methods. This is fully in line with earlier propositions from the practical environmental decision analysis literature (Gregory et al., 2012, p.212). Very similarly, the literature on biases in MCDA proposes the use of multiple methods to help overcome well-known biases such as range insensitivity during weight elicitation (Montibeller and Winterfeldt, 2015).

There are different ways to evaluate methods. The strict experimental approach stems from the natural or e.g. economic sciences. Hereby large sample sizes (N) are needed along with an experimental setup in a controlled environment to allow for statistical testing of the research hypotheses. Such a strongly formalized analysis restricts the dimensions of the problem. In the literature, one usually finds constructed or smaller research problems, and test persons are often students. The second possible approach to compare methods is more qualitative, but not necessarily less relevant. It focuses on real-world interventions, on real decision makers, and on the needs of practitioners. Real-world decisions are usually complex, involving many objectives, complex strategic alternatives, and many stakeholders that pursue different interests. For example, our meta-analysis of 61 environmental and energy MCDA application cases revealed an average of 15 objectives, ranging from 3 to 51 (Marttunen et al., 2017). Thus, didactical examples with some five objectives can provide insights, but do not mirror the challenges encountered in real cases. The literature increasingly emphasizes the need for such real-world applications, even if sacrifices regarding hypothesis testing need to be done (see the recent BOR literature; e.g. Franco and Hämäläinen, 2016).

The aim of this study is to register the implementation impacts of two different MCDA philosophies (aggregation and disaggregation) in a very complex real and typical environmental decision with ten stakeholders through two series of interviews concerning sustainable wastewater infrastructure planning (SWIP). Rather than pursuing rigorous experimental testing, we aim to empirically learn about the advantages and disadvantages of the two paradigms, hereby also addressing the needs of MCDA practitioners. Nevertheless, we tried to be as rigorous as possible in our application; and consider ourselves as exceptionally lucky that we were able to interview the same ten stakeholders twice – once with each approach. The motivation for using different elicitation methods is thus to avoid the systematic bias a specific method may have and to enhance the trustworthiness of the final result. We investigated the following issues:

1. How to design a practical elicitation procedure for indirect disaggregation methods for a complex decision context? The environmental case study involves a large objectives hierarchy with 19 lowest-level fundamental objectives. This could possibly require stakeholders to compare a large number of reference alternatives two-by-two, but typically stakeholders have limited time for the elicitation task.
2. Is the elicited MAVT model consistent for the two different elicitation methods?
3. Is the evaluation of alternatives using elicited preferences obtained by

the two different elicitation methods consistent?

4. How do the stakeholders and the analyst perceive the two elicitation methods in the application; i.e. what are the perceived differences between methods?

The remaining paper is organized as follows: Section 2 describes the case study and the general design of the elicitation interviews. Section 3 presents the preference elicitation methods and interview procedures. Section 4 presents the elicited preference parameters (weights, value functions) and evaluates six real-world alternatives. The results of the two different interviews are compared and inconsistencies are analyzed. Section 5 discusses interesting findings. The paper concludes in Section 6.

2. Case study

The wastewater infrastructure system provides multiple benefits to society. Originally put in place to grant urban hygiene, it is of crucial importance to human health. Wastewater systems also provide environmental benefits by protecting fresh water resources from pollutants. However, the centralized wastewater system is increasingly criticized for sustainability reasons and because it relies on massive infrastructure networks: the sewer pipes. This system is highly inflexible, ageing, and expensive (reviews see e.g. Gleick, 2003; Larsen et al., 2016). The required global investments in water infrastructures are estimated to exceed 500 billion U.S.$ per year (Milly et al., 2008). The water infrastructures are very long-lived. Sewers have lifetimes of around 80 years, which makes planning extremely uncertain. The uncertainty increases due to climate change effects, droughts, and heavy rainfalls leading to sewer overflows, and other future changes such as population growth. However, current infrastructure planning is based on mid-term projections of the status quo and on only few objectives, even though many fields are impacted by water infrastructures. Moreover, despite affecting the population, the decisions are usually not participatory.

The transdisciplinary SWIP project (Sustainable water infrastructure planning[1]) was set up to address these challenges in an exemplary Swiss case study. There was close collaboration between urban water engineers and decision analysts. We proposed MCDA to make better informed, more sustainable and participatory decisions about wastewater infrastructures.

[1] http://www.eawag.ch/en/department/sww/projects/sustainable-water-infrastructure-planning-swip/

Four municipalities were involved in the case study near Zürich, totaling about 24'000 inhabitants in 2010. There are three wastewater treatment plants, which are, however, reaching their capacity limits. The cleaned wastewater eventually flows into Lake Greifensee, which is impacted by too high nutrient levels (Känel et al., 2008). The sewers have an average age of 33 years; and investment decisions need to be made in the coming years.

These investment decisions involve high stakes and have long-term consequences on social, environmental, and economic aspects. For the research presented here, we conducted two preference elicitation interviews in 2013 with each stakeholder individually without any possibility of interaction between stakeholders. We decided against group decision-making mainly because we were interested in the perspectives of different interest groups. Group decisions risk that individual views get lost. We carefully selected our ten interview partners with a stakeholder and social network analysis, based on 27 preliminary semi-structured qualitative interviews in 2010–11 (Lienert et al., 2013). The stakeholders were representatives of different decisional levels (local, cantonal, and national), they pursued different interests, and came from different sectors (engineering practice, administration and politics, science; Zheng et al., 2016). Another advantage of individual interviews was that they allowed us to focus on methodological aspects. It would have been very difficult to elicit the entire set of preference parameters (weights, marginal value functions) from a group consensus process. Additionally applying the pairwise comparisons for the disaggregation approach would have simply been impossible in a group.

However, we found it important that the stakeholders share the same decision framework. Therefore, the objectives hierarchy and attributes, the decision alternatives, and four future scenarios used in the entire SWIP project (i.e. also for water supply) were set up in group workshops in 2011 (see in-depth description of process in Lienert et al., 2015).

The stakeholders were thus acquainted with the decision context before we elicited their preferences concerning the objectives hierarchy in Figure 1. Five top-level objectives were judged as fundamental to achieve our goal: *Sustainable wastewater infrastructure* (Sustainable WWI), including *Intergenerational equity* (Equity), *Protection of water and other resources* (Protection), *Safe wastewater disposal* (Safe WW disposal), *High social acceptance*, and *Low costs.*

At the end of the project, we carried out a large stakeholder event in 2014, where the 100 participants - including the stakeholders interviewed in this project – could discuss their opinions and experiences.

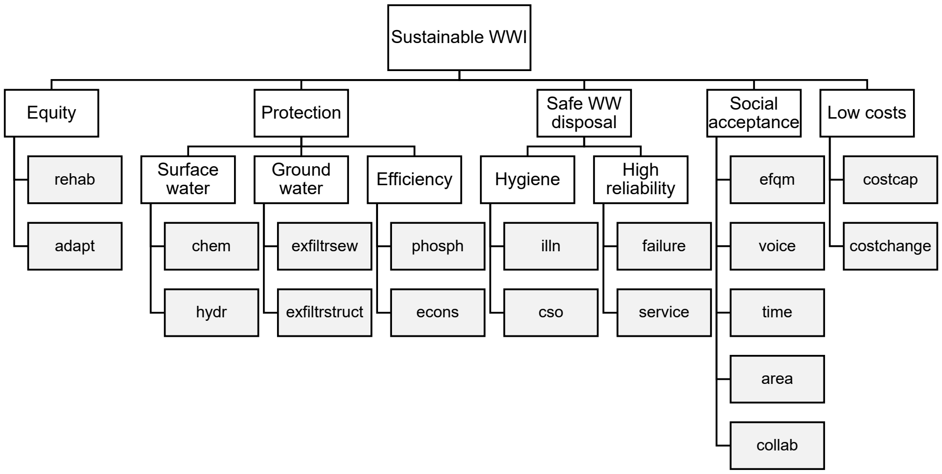

Figure 1: Objectives hierarchy of wastewater infrastructure planning. The five top-level main objectives are split into lower-level sub-objectives. The lowest-level sub-objectives (the shaded boxes at the bottom, see Table A.1 in Appendix A for their meanings) are measured by attributes with identical names (see Table SM-1 in Supplementary Material for definitions). WWI=Wastewater infrastructure; WW=Wastewater.

3. Methods

3.1. MCDA model

We used the widely applied additive value model where the overall value of alternative a is the weighted average value of all attributes:

$$v(a) = \sum_{i=1}^{n} w_i v_i(a) \tag{1}$$

$v_i(a)$ is a marginal value function normalized on [0,1], reflecting the value of alternative a on attribute i. w_i is a scaling constant (=weight) of the ith attribute, indicating the importance of improving attribute i from the worst to the best case. Weights should be non-negative and sum up to one.

3.2. Direct aggregation methods for preference elicitation

3.2.1. Elicitation procedure and preference modeling

In a parallel research project about drinking water infrastructure planning (e.g. Scholten et al., 2015), we had used the standard SWING method to elicit preferences. These experiences with real stakeholders (from the water supply sector) raised some issues, which we aimed to improve. In this wastewater case, we therefore adapted SWING to the SMART/SWING-variant. To elicit marginal value functions, we used the popular bi-section method. This fairly standard elicitation procedure has been described in our earlier publication (Zheng et al., 2016), so we only briefly recall the main points here.

Two weeks before the interview, the stakeholders received a package per post containing an information letter describing the project, the objectives hierarchy, and the attributes including their ranges (best and worst possible cases). They were asked to answer an online questionnaire for weight elicitation in which we had implemented a variant of the interval SMART/SWING method (Mustajoki et al., 2005; Zheng et al., 2016). Firstly, the respondents chose a reference objective, with which they were most familiar with. Then they made pairwise comparisons of the other objectives with the chosen reference. Finally, they stated their strength of preference by choosing from nine categories (as in AHP; Saaty, 1980). Online elicitation was used to familiarize the stakeholders with the topic and elicitation method.

The online questionnaire was followed by a personal interview carried out by a psychologist with the support of an MCDA analyst, again using the SMART/SWING-variant. We present an illustrative example of the elicitation tools with more details and visualization for the sub-objective *Intergenerational equity* in Section SM-2.1 of the supplementary material. The weights were calculated by interpreting the indicated strength of preference as ratio between the weight of the reference objective and the weight of another objective (see Zheng et al., 2016).

For the most important attribute identified from weight elicitation, we elicited the marginal value function in detail with the common bi-section method (Eisenführ et al., 2010). Due to time constraints, only rough elicitation was performed for the other attributes, asking: “Which is more important to you, improving the objective from its worst-possible case to the mid-point evaluation level, or improving the objective from the mid-point evaluation level to its best-possible case?” This is a simplified way to only elicit minimum information about the shape (concave, convex, or linear) that saves a lot of time (Scholten et al., 2015; Zheng et al., 2016). The points elicited with the bi-section method were connected by piecewise functions. For rough elicitation, exponential functions were assumed (see Zheng et al., 2016, for details).

3.3. Indirect disaggregation methods for preference elicitation

Additionally, for MAVT from an elicitation point of view and again based on the experiences with SWING, we were interested to find out, whether it would be helpful for stakeholders to infer preferential parameters by using indirect preference information, rather than eliciting the preference parameters directly (as suggested by e.g. Greco et al., 2008; Kadziński et al., 2016; Kadziński and Tervonen, 2013; van Valkenhoef and Tervonen, 2016). To be as consistent as possible, we aimed to stay within the MAVT paradigm, thus excluding e.g. outranking approaches. For this indirect disaggregation elicitation framework, we applied the UTA$^{\text{GMS}}$ method as it satisfied our requirements in the case study. UTA$^{\text{GMS}}$ applies pairwise comparisons, which are claimed to mimic intuitive decision behavior, but is less strongly criticized than AHP (e.g. Macharis et al., 2004; Smith and von Winterfeldt, 2004). A main argument for our choice of UTA$^{\text{GMS}}$ is that it can be used interactively with stakeholders. This is very important for this typical environmental case, where we had 19 lowest-level objectives (see Fig. 1). Asking for pairwise comparisons could have resulted in a ridiculously large number of questions. Thus, we aimed to find a method that allowed us to design an interactive procedure. Hereby, it became possible to reduce the number of questions by inferring necessary relations from previous answers, instead of asking questions for these. This elicitation procedure allows the stakeholders to progressively express their preference information so that the necessary relation is enriched. We emphasize once more that the stakeholders were willing to carry out preference elicitation twice; just for research reasons, and openly shared their experiences with each method. We consider this as a very unique opportunity to be able to empirically compare within-individual preferences of important stakeholders derived with two different methodological approaches.

3.3.1. Existing elicitation methods

The number of questions is influenced by the selection and ordering of questions. This issue has been addressed in different fields. In conjoint analysis, adaptive methods have been developed to collect more information per question; using far fewer questions than traditional methods (e.g. Toubia et al., 2004). In electronic commerce, there are often tasks of evaluating and ranking items with multiple attributes. A heuristic has been proposed for choosing pairwise comparisons based on the idea of weight space so that a preference model can be built with only a small number of user queries (Iyengar et al., 2001). In the field of MCDA, researchers recently apply heuristic approaches for selecting pairwise elicitation questions in an interactive process for either choice or ranking problems (Branke et al., 2017; Ciomek et al., 2017a,b). These methods focus on reducing the uncertainty of decisions measured by different metrics, such as information gain (van

Valkenhoef and Tervonen, 2016), the number of alternative pairs for which the necessary preference relations holds, etc (Ciomek et al., 2017a). Results of the experimental studies indicate that the best performing heuristics in terms of minimizing the number of questions depend on the specific measure of uncertainty.

3.3.2. Elicitation method

We designed an elicitation procedure based on UTA$^{\mathrm{GMS}}$ to elicit the preferences with pairwise comparisons (Greco et al., 2008), for the reasons stated above. If the preference information was consistent, i.e. the stated preferences could be represented by an additive value function, mostly there existed multiple additive value functions compatible with the constraints derived from such information. In rare cases, a unique value function could be determined.

We would like to point out that additive value functions do not necessarily represent the preferences of stakeholders. This can be accounted for by using other aggregation models for MAVT (see Langhans et al., 2014; Reichert et al., 2015). For example, Cobb-Douglas aggregation is able to model the veto power of an attribute, i.e. that a zero value of one attribute leads to an overall zero value of the assessment. In our wastewater application case, not all stakeholder preferences were well represented by the additive model because, for instance, the strong preferential independence assumptions did not always hold. However, with systematic sensitivity analyses we were able to show that the ranking of alternatives was in most cases only slightly influenced if a different aggregation model was considered (Zheng et al., 2016). As it requires substantial additional effort to elicit a more expressive aggregation model, we regard it as justified to assume the additive aggregation model in this case study.

Moreover, we point out that our application involves a hierarchy of objectives. We used UTA$^{\mathrm{GMS}}$ at each level of the hierarchy to obtain the preference relation with respect to the set of criteria at the same level of the hierarchy. The Multiple Criteria Hierarchy Process (MCHP) in ROR permits consideration of preference relations with respect to a subset of criteria at any level of the hierarchy (Corrente et al., 2012). However, for this application it is not necessary to consider the preference relation while considering a subset of criteria at any level of the hierarchy, which is why we did not use MCHP.

Two key concepts of the UTA$^{\mathrm{GMS}}$ method were defined:

1. necessary weak preference relation: alternative a is **necessarily** ranked

as at least as good as b if and only if $v(a) \geqslant v(b)$ for **all** value functions compatible with the preference information;

2. possible weak preference relation: alternative a is **possibly** ranked as at least as good as b if and only if $v(a) \geqslant v(b)$ for **at least one** value function compatible with the preference information.

When no preference information was available, the necessary weak preference was merely a weak dominance relation, i.e. alternative a was preferred to alternative b when a was at least as good as b regarding all attributes. At this point the possible weak preference was complete. When additional preference information was provided, the necessary weak preference was enriched while the possible weak preference was impoverished.

3.3.3. Interview procedure

Approximately three months after the first interview, the UTA$^{\text{GMS}}$ interview was conducted by the same interviewers with the same ten stakeholders. No preparation was required this time.

The UTA$^{\text{GMS}}$ method was designed to be used interactively, i.e. a computer-aided exchange between analysts and stakeholders. The program progressively added each piece of preference information and selected the next pairwise comparison, which we describe later. Elicitation was carried out at each level of the objectives hierarchy, and the elicitation was only stopped after we obtained a complete ranking of the reference alternatives. In other words: we used UTA$^{\text{GMS}}$ interactively to control for the robustness of the MAVT model. We applied UTA$^{\text{GMS}}$ hierarchically in a bottom-up way. For the pairwise comparisons, we did not explicitly use “real” alternatives but designed hypothetical reference alternatives. These were created in different manners depending on the specific sub-objectives of the hierarchy (Fig. 1).

Objectives with two lowest-level sub-objectives In this case, e.g. for the main objective *Intergenerational equity*, the marginal value functions and weights of attributes (e.g. “rehab” and “adapt”; Fig. 1) had to be determined. We used hypothetical alternatives designed by a three-level full factorial plan. These levels corresponded to the best, mid-point, and worst case of the attribute. There were $3 \times 3 = 9$ such alternatives to be ranked (Tab. 1). After 27 weak dominance relations were eliminated, the stakeholder only needed to compare 9 pairs to determine the ranking of the 9 reference alternatives Comp$_1$: $C(a_3, a_7)$, Comp$_2$: $C(a_2, a_7)$, Comp$_3$: $C(a_2, a_4)$, Comp$_4$: $C(a_3, a_4)$, Comp$_5$: $C(a_3, a_5)$, Comp$_6$: $C(a_3, a_8)$, Comp$_7$: $C(a_5, a_7)$, Comp$_8$: $C(a_6, a_7)$, Comp$_9$: $C(a_6, a_8)$. The comparisons were directly presented to the stakeholders, with emoticons and pictures depicting the different cases. The preference (indifference, resp.) statement that alternative a_i is preferred to a_j (a_i is indifferent to a_j, resp.) is denoted as $a_i \succ a_j$ ($a_i \sim a_j$).

Table 1: Reference alternatives for eliciting preferences for the objective *Intergenerational equity* with two sub-objectives “rehab” and “adapt”. Green smiley is the best case, yellow is the mid-point, and red is the worst.

	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9
Rehabilitation	🙁	😐	🙂	🙁	😐	🙂	🙁	😐	🙂
Flexibility	🙁	🙁	🙁	😐	😐	😐	🙂	🙂	🙂

The elicitation started by comparing two extreme alternatives a_3 and a_7. The stakeholder could respond with a preference or indifference statement (Box 3, SM). We then asked her for her strength of preference (Box 4, SM): “How much is the difference of attractiveness between the two alternatives?” by choosing one of the nine above-mentioned categories: C_1= no difference; C_3= weak difference; C_5= moderate difference; C_7= strong difference; C_9= extreme difference; C_2, C_4, C_6, C_8= between categories. Classification of the statement $a_i \succ a_j$ into one of the nine categories was denoted as $C(a_i \succ a_j) \in C_k$. To enrich the preference information, a trade-off question followed if the stakeholder did not make an indifference statement: she was asked to worsen the preferred alternative by deteriorating the objective in the best state so that the two alternatives were equally good (Box 5, SM).

It was unnecessary to compare all the remaining eight pairs, because the relation of some pairs became a necessary weak preference relation with the first preference statement. With each new answer, the necessary and possible weak preferences were computed (see section 3.3.4). The next question was chosen by the computer program based on the strategy that we wanted to minimize the number of questions. For each possible weak preference relation (e.g. c and d), we again computed the number of necessary weak preference relations $n_{\succ}$ ($n_{\sim}$, or $n_{\prec}$, respectively) assuming that $c \succ d$ (or $c \sim d$, $c \prec d$, respectively). The more new necessary weak preference relations this possible preference relation generated ($n_{\succ} + n_{\sim} + n_{\prec}$), the more informative this question was. Then one comparison which would generate the most necessary weak preference relation was chosen. The process was terminated when the necessary weak preference was complete in the reference alternative set as shown in Table 1.

An illustrative example of the interaction process for the sub-objectives of *Intergenerational equity* is given in SM, Section-2.2.1.

Objectives with more than two lowest-level sub-objectives This only concerns the main objective *Social acceptance* which consists of five sub-objectives (Fig. 1). We developed hypothetical alternatives where each sub-objective could have two levels. To speed up elicitation, we only asked for a ranking of the five hypothetical alternatives, which each had one sub-objective on the best level and the others on the worst levels. If *Social acceptance* was later stated important (considering the ranges) in the elicitation process of the five main objectives (see below), we went back and carried out further elicitation. Hereby, one sub-objective was chosen as reference with which the others were compared (same procedure as "objectives with two lowest-level sub-objectives"). The strength of preference and trade-off questions were asked as well.

Higher-level sub-objectives For objectives at a higher hierarchical level, only weights had to be derived, but no marginal value functions. Hypothetical alternatives with sub-objectives having two levels (best and worst cases) were designed. For the five main objectives, the fractional factorial design 2^{5-1} built 16 reference alternatives which had two levels for each objective. We first asked the stakeholder to rank the five alternatives (one objective on best, all others on worst level), as in the standard SWING method. From the responses, some necessary weak preference relations were derived and more pairs of possible weak preference relations were interactively compared. For the sub-objective *Protection* (*Safe WW disposal*, resp.), a full factorial design was used to construct hypothetical alternatives. This resulted in $2^3 = 8$ ($2^2 = 4$, resp.) reference alternatives to be compared for each. It is easy to verify that for *Protection* (*Safe WW disposal*, resp.) there was only six (one, resp.) pair of alternatives to compare after dominance relations were eliminated. We used the same strength of preference and trade-offs questions as above.

3.3.4. Preference modeling

The weights and marginal value functions were inferred by ordinal regression via linear programming. The inferred MAVT model restores the pairwise comparisons provided by the stakeholders. The inference was performed at each level of the objectives hierarchy.

Let us introduce some notations as in Greco et al. (2008). A set of reference alternatives $A^R = \{a_1, a_2, ..., a_j, ..., a_m\}$ is evaluated on n criteria $g_1, g_2, ..., g_i, ..., g_n$ ($i \in G = \{1, ..., n\}$). The evaluation scale on criterion g_i

is X_i, i.e. $g_j : A \mapsto X_i$. We assume that the evaluation scale is bounded, i.e. $X_i = [\alpha_i, \beta_i]$ where α_i and β_i are the worst and the best evaluations. Therefore, each alternative a can be represented by a profile $g_1(a), ..., g_n(a)$ in the evaluation space $X = \prod_{i \in G} X_i$. We use $v_i(a)$ to replace $v_i(g_i(a))$ to represent the value function of attribute i. Let us denote the permutation on the set of indices of alternatives from A^R that reorders them according to the increasing evaluation on g_i as τ_i, i.e. $g_i(a_{\tau_i}(1)) \leq g_i(a_{\tau_i}(2)) \leq \cdots \leq g_i(a_{\tau_i}(m-1)) \leq g_i(a_{\tau_i}(m))$. These are called characteristic points of the marginal value function $v_i(\cdot)$.

The linear constraints and objective function of the linear programming are:

Constraints

The stakeholder stated that a was preferred to b and classified this into one of the nine categories, i.e., $C(a \succ b) \in C_k$. She also classified her preference of c to d to a category $C(c \succ d) \in C_{k'}$. Suppose that $k > k'$. The pairwise comparisons were transformed into linear constraints:

$$\begin{cases} v(a) - v(b) \geq v(c) - v(d) + \epsilon \\ v_i(a_{\tau_i}(j)) - v_i(a_{\tau_i}(j-1)) \geq 0 \quad \forall i \in G; j = \{2, \cdots, m\}. \\ v_i(a_{\tau_i}(1)) \geq 0; \quad v_i(\beta_i) \geq v_i(a_{\tau_i}(j)) \quad \forall i \in G; j = \{2, \cdots, m\}. \\ v_i(\alpha_i) = 0, \quad \forall i \in G \\ \sum_{i=1}^{n} v_i(\beta_i) = 1 \end{cases} \tag{2}$$

The values of the alternatives can be calculated additively, from the characteristic points of the evaluation criteria, e.g. $v(a) = \sum_{i=1}^{n} v_i(a)$. The other constraints guaranteed that the value functions were monotonic and the value of the best (worst, resp.) alternative was 1 (0, resp.). The conditions of program (2) were all transformed into linear constraints. We did not make any assumption about piecewise linearity of the marginal value functions. The formulation of program (2) modeling the strength of preferences has been proposed in earlier research (Bana et al., 2016; Figueira et al., 2009; Hurson and Siskos, 2014), and we adopted the semantic categories of preference intensity in AHP to support the stakeholders to make judgments (Saaty, 1980).

Objectives

First, the necessary and possible weak dominance relations had to be computed to support interactive elicitation. To determine the comparison of alternative e and f, two linear programming problems need to be solved by maximizing $v(e) - v(f)$ and $v(f) - v(e)$, subjecting to the constraints (2).

The relation of e and f could be determined from the two maximum values according to the rules described in Greco et al. (2008).

Second, we solved a different linear programming problem to identify one representative additive value function model to evaluate the "real" alternatives by maximizing ϵ.

At each level, the weight of one objective was obtained as the value of the best evaluation: $w_i = v_i(\beta_i)$. In other words, the value functions can be normalized by weights as in (1).

3.4. Determining the consistency of the results between the two interviews

The comparison of the elicitation results of the two interviews concerned several aspects:

Weights. For the SMART/SWING-variant, the calculation of weights assumed that the semantic concerning the stated strength of preference can be interpreted as weight ratio. For UTA$^{\mathrm{GMS}}$, the weights were inferred so that they were compatible with some pairwise comparisons. Hereby, the strength of preference statement was interpreted as value differences between the reference alternatives. However, choosing one additive value function from a set of compatible models is somewhat arbitrary. Thus, changes in weights resulted from both the changes in preferences and different modeling techniques. Therefore, we calculated the Euclidean distance to measure the weight differences of the two interviews, but paid more attention to analyze the changes in the rankings of the weights. We used the following indicators: 1) Kendall's τ rank correlation coefficient between the two sets of weights (Kendall, 1938). To deal with equal values (ties), we followed Amerise et al. (2015); 2) we calculated the proportion of rank reversals of all pairwise comparisons, concerning ranking the importance of (sub-)objectives (Fig. 1); 3) we also calculated the proportion of cases where the indifference statement was changed to a preference relation out of all pairwise comparisons; and 4) the proportion of changes of the preference statement to an indifference relation.

Marginal value functions. In the first interview, we only had knowledge about the rough shape of the marginal value functions for most attributes. For the interview using UTA$^{\mathrm{GMS}}$, there were multiple compatible value functions, and sometimes the constraints from the preference information were sufficiently loose to allow for both concave or convex functions. Therefore, we did not statistically compare the elicited marginal value functions because of insufficient information.

MCDA evaluation results.

The elicited additive value functions were used to evaluate the performance of six wastewater infrastructure alternatives for the case study: A2 (central high-tech system), A5 (decentral low-tech system), A7 (decentral system with nutrient recovery), A8a (central system with stormwater retention), A8b (decentral high-tech system), and A9 (central privatized system). The alternatives were selected as the most typical and most strongly discriminating in the SWIP-project and the labels were kept as in Zheng et al. (2016). A detailed definition of these alternatives can be found therein. The outcomes of the alternatives on the 19 attributes are given in Table SM-1.

The overall values and rankings of alternatives for the two interviews were compared: 1) by asking whether the elicited preferences in the two interviews lead to a congruent best alternative; 2) by calculating Kendall's τ rank correlation coefficient between the two rankings; (3) by calculating the Euclidean distance of the two overall values of alternatives.

Stakeholder feedback. At the end of the second interview, we calculated the weights based on the stakeholders' responses. If answers were inconsistent, i.e. the rankings of the importance of objectives were different compared to the interview using the SMART/SWING-variant, we explicitly asked the stakeholders for possible reasons. The stakeholders were also asked which of the two interviews they perceived to be more difficult. Any other remarks and comments were encouraged and recorded.

3.5. Implementation and analyses

We used R for all analyses and statistical tests (R Development Core Team, 2017). The online questionnaire to prepare the interviews and give a preliminary assessment of the weights was implemented in a trial version of Qualtrics (`https://www.qualtrics.com`). The interactive UTA$^{\text{GMS}}$ procedure that assisted the personal interviews was also implemented in R using the *ror* package (R Development Core Team, 2017; Tervonen, 2013). To calculate the additive value model, we used the R package *utility* (Reichert et al., 2013).

4. Results

We analyzed the average weights of each of the ten stakeholders in the two interview series. The answers from the online questionnaire were not used, because 1) the online questionnaire was designed only for preparing the interview; 2) three stakeholders did not answer it; and 3) for those who responded, the answers were sometimes not complete as we had provided opt-out options.

4.1. Individual preference changes

For individual stakeholders there were some preference changes in the second interview using UTA^{GMS}, compared to the first using the SMART/SWING-variant for weight elicitation. The complete data of the weights and their rankings is given in the supplementary material (Tabs. SM-3, SM-4). At the level of main objectives, the changes in the weights were somewhat larger for SH1–SH3 than for the other stakeholders (weight points are further away from the diagonal: Fig. 2); this is also reflected in the larger Euclidean distances (Tab. 2).

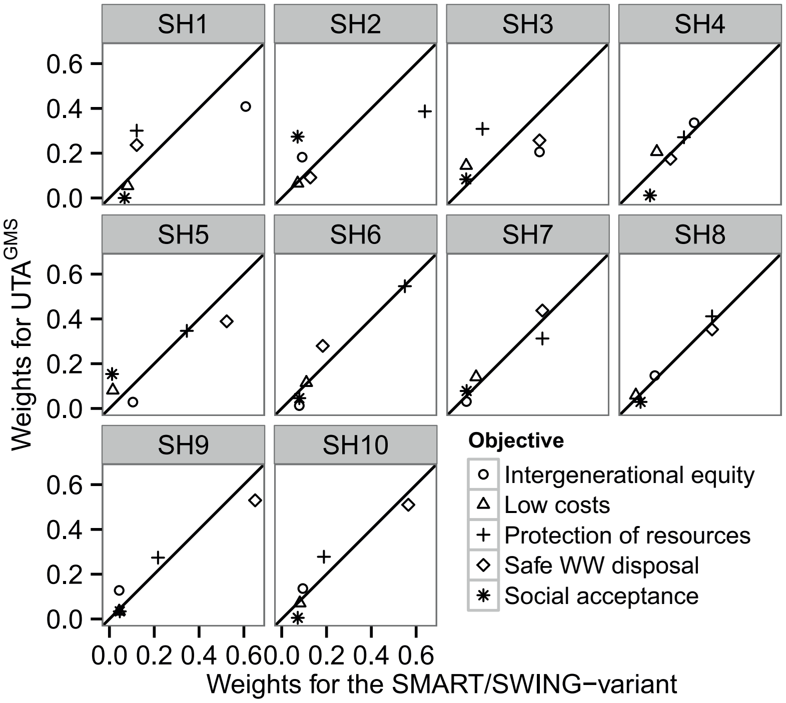

Figure 2: Weights of main objectives for ten stakeholders (SH) elicited in two interviews. x-axis: weights obtained with the SMART/SWING-variant; y-axis: weights obtained with UTA^{GMS}. Objectives that received the same weights in both interviews appear on the diagonal line. WW=Wastewater.

For all ten stakeholders, the rankings of the five main objectives were positively correlated between the two interviews (Tab. 2). The ranking of the five main objectives between the interview using the SMART/SWING-variant and using UTA$^{\text{GMS}}$ did not at all change for one stakeholder (SH10; Kendall's τ= 1). For five stakeholders (SH1, SH4, SH6–SH9), the rankings were relatively strongly correlated, between 0.6 and 0.9 (Tab. 2). Three stakeholders (SH2, SH3, SH5) had relatively large changes in the rankings, resulting in low correlations between 0.3 and 0.4. On average over all stakeholders, 11% (±11%) of the importance rankings of main objectives were reversed. For another 9% (±7%) of the comparisons, the stakeholders changed from indifference to preference statements. The case that preference relations were changed to indifferent relations rarely happened (only SH9 for one comparison).

Table 2: Changes in weights of the main and lowest level objectives in the two interviews with ten stakeholders (SH) using the SMART/SWING-variant and UTA$^{\text{GMS}}$. For each stakeholder, we present Kendall's τ correlation coefficient (*Cor*) of the two weight rankings, the proportion of rank reversals (*RR*), the proportion of changes from an indifference to a preference statement (*IP*), the proportion of changes from a preference to an indifference statement (*PI*); and the Euclidean distance of the two weights (*d*). Last rows: average (*Ave.*) and standard deviation (*SD*) for all stakeholders.

	Main objectives					Lowest level sub-objectives				
	Cor	RR	IP	PI	d	Cor	RR	IP	PI	d
SH1	0.9	0%	10%	0%	0.30	0.35	16.1%	6.5%	0%	0.40
SH2	0.3	30%	10%	0%	0.34	0.39	6.5%	22.6%	0%	0.27
SH3	0.4	20%	20%	0%	0.30	0.54	9.7%	35.5%	0%	0.28
SH4	0.8	10%	0%	0%	0.12	0.40	29.0%	3.2%	0%	0.19
SH5	0.4	30%	0%	0%	0.22	0.23	16.1%	12.9%	3.2%	0.22
SH6	0.9	0%	10%	0%	0.12	0.42	3.2%	25.8%	6.5%	0.24
SH7	0.8	0%	20%	0%	0.11	0.61	3.2%	22.6%	0%	0.14
SH8	0.7	10%	10%	0%	0.06	0.59	12.9%	9.7%	9.7%	0.14
SH9	0.6	10%	10%	10%	0.16	0.61	3.2%	9.7%	0%	0.16
SH10	1	0%	0%	0%	0.13	0.60	6.5%	12.9%	0%	0.16
Ave.	0.68	11%	9%	1%	0.19	0.47	10.6%	16.1%	1.9%	0.22
SD	0.24	11.4%	7.0%	3.0%	0.09	0.13	7.8%	9.6%	3.3%	0.08

It is worth noting that a stable ranking of the objectives did not necessarily lead to similar weights because the two indicators (ranking measured by Kendall's τ correlation/ weights measured as Euclidean distances) do not reflect the same thing. For example, SH1 had a very stable ranking of the weights of the main objectives (Kendall's τ = 0.9 and 0% rank reversals; see Tab. 2), but the weights differed relatively strongly between the first and second interview (relatively large Euclidean distance d of 0.3; Tab. 2). Likewise, strong changes in the ranking did not necessarily result in dramatic changes in weights. SH9 serves as an example: the correlation of the ranking of the weights between the two interviews was only 0.6 for main objectives, and he had 10% rank reversals (Tab. 2), but the weights were relatively similar (near diagonal; Fig. 2), which is reflected in a low Euclidean distance of 0.16 (Tab. 2).

The ranking of the 19 lowest level sub-objectives were indirectly derived from their global weights, which were hierarchically calculated from the local weights of the lowest level sub-objectives and the associated higher level objectives. Therefore, weight changes of these 19 sub-objectives were a result of weight changes of objectives at all hierarchy levels. As for the highest hierarchy level, the changes in weights at the lowest level were relatively larger for SH1–SH3 than for the other stakeholders (Tab. 2; larger Euclidean distances). Nevertheless, we found a positive correlation between the two rankings for all stakeholders, although Kendall's τ was generally relatively low for the sub-objectives. At lowest hierarchy levels, 17 pairs of these sub-objectives had to be compared (Fig. 1). Also for these lowest level sub-objectives, rank reversals occurred (10.6% ± 7.8%), as well as changes from indifference to preference statements (16.1% ± 9.6%). However, preference relations very rarely changed to indifference statements (1.9% ± 3.3%).

4.2. Comparison of MCDA results of two interviews

For six stakeholders, the best-performing alternative (i.e. rank 1) was the same in the first and second interview (but not for SH1, SH7, SH8, and SH10; Tab. 3) . For seven stakeholders (except SH6, SH9, SH10), the rankings of all six alternatives were significantly correlated in a positive direction between the first and second interview (all Kendall' $\tau > 0.7$, $p < 0.05$, Tab. 3; marginally significant for SH10). Particularly, the complete rankings were kept for SH2 and SH4. For SH6, SH9, and SH10, the two rankings were positively correlated, but the correlations were not statistically significant. An extreme case was SH9, where there was almost no correlation between the two rankings. The overall values of alternatives changed between the first and second interviews for different stakeholders to different extents (see row d; Tab. 3). We discuss possible reasons for the changes for two extreme cases, namely stakeholders SH1 with the highest value change, and SH9 with the greatest rank change (for reasons of space see Supplementary material; SM Section 4, Tabs. SM-6–SM-7, Figs. SM-3–SM-4).

Table 3: Evaluation results of six wastewater alternatives for ten stakeholders (SH) in the two interviews (SMART/SWING-variant for weights and bi-section method or rough elicitation to determine shape of value functions; UTA$^{\mathrm{GMS}}$). The analysis was made for each stakeholder (SH1–SH10). We show whether the best-performing alternative (rank 1) is congruent for the first and second interview (*Con*, Y = Yes, N = No). *Cor*: Kendall's τ correlation coefficient between the rankings of all alternatives for each interview; *P-value*: under null hypothesis of a non-positive correlation; *d*: Euclidean distance of the two overall values of alternatives.

	SH1	SH2	SH3	SH4	SH5	SH6	SH7	SH8	SH9	SH10
Con	N	Y	Y	Y	Y	Y	N	N	Y	N
Cor	0.73	1	0.87	1	0.73	0.47	0.73	0.73	0.07	0.6
P-value	0.03	0	0.01	0	0.03	0.14	0.03	0.03	0.5	0.07
d	0.5	0.23	0.24	0.32	0.06	0.08	0.06	0.08	0.22	0.05

4.3. Feedback from stakeholders

4.3.1. Reasons for preference instability

We summarized the explanations of the stakeholders for their changes of preferences in the second interview compared to the first, and categorized them into five classes (complete documentation in Tab. SM-8). Note that the classification is subjective and that the five classes may overlap. For each class, a few examples are presented for illustration.

Different decision strategy: In 16 cases, the stakeholders attributed their preference change to a change in perspective. For instance, when comparing the objective *Few structural failures of drainage system* and *Sufficient drainage capacity of drainage system*, stakeholder SH5 said he paid more attention to the negative consequences on human beings in the second interview, and therefore changed his mind from a preference (*Structural failure* being more important) to an indifference relationship.

Uncertainty of preferences: Some stakeholders explicitly stated in nine cases that their preferences changed, simply because they were not sure about them. There might be different reasons for this uncertainty. For instance, stakeholder SH3 emphasized the difficulty of comparing the objectives *Low future rehabilitation burden* and *Flexible system adaptation* because he found that the two objectives were related to each other. Sometimes the long-term consequences of the wastewater infrastructure on the environment were unclear, making the comparison difficult. Stakeholder SH10 pointed out that she was unsure about the consequence of biocides on groundwater.

Different elicitation methods: The stakeholders mentioned in eight cases that the different formulations of the questions in the two elicitation methods may have had an effect on the answers. For instance, stakeholder SH7 ranked *High social acceptance* higher than *Intergenerational equity* in the second interview rather judging them indifferent (first interview). He thought that the questions in the second interview were a good way to reach a decision faster. The different perception of the two elicitation methods are presented in more detail below (Section 4.3.2).

Learning effect: The stakeholders mentioned seven times that they had gained more insight into the decision problem between interviews, by being able to reflect more on the decision, by again discussing with the analyst, or simply by answering our questions. As an example, stakeholders SH2, SH3, SH4, and SH5 thought that their changes of preferences for the main objectives could to some extent be attributed to such a learning effect when working through the topic twice.

External influence: The stakeholders mentioned six times that their preferences were influenced by external events in between interviews, or that they had gained new insights by obtaining new information. As an example, stakeholder SH1 stated in the second interview that the sub-objective *Fewer structural failures of drainage system* was more important than *Sufficient drainage capacity of drainage system* instead of an opposite preference in the first. He explained that recently there had been a pipe failure in his neighborhood, which had caused a traffic chaos. He had also learned that the current system in Switzerland is designed for very bad cases and that the system is therefore securely over-dimensioned.

4.3.2. Comparison of methods by stakeholders

The stakeholders were asked to compare the two interviews, after the different weights of the main objectives in the two interviews were presented to them and after discussing the inconsistencies between the two rankings of objectives/ sub-objectives (details see Tab. SM-9). The main message is:

Difficulty: For six of the ten stakeholders (SH1, SH3-5, SH7, and SH10), the elicitation methods in the second interview were easier than the ones in the first. The questions in the second interview were considered as more direct (SH4) and understandable (SH3). Also familiarity with the topic after working through it a second time might be the reason that the second interview was easier (SH3, SH5, and SH7). Especially the elicitation of marginal value functions in the first interview was perceived as difficult (SH10). Only SH6 judged the first interview to be easier because the cognitive load was lower. For SH8, the first interview was methodologically more difficult, i.e. the questions were difficult, but the second interview was more demanding when having to decide between the hypothetical alternatives. The difficulty of having to make trade-offs was acknowledged by several stakeholders for both interviews (SH1, SH2, SH10). SH2 and SH9 did not see clear differences between the interviews.

Features: SH6 stated that there was less cognitive load to answer the questions in the first interview and that the answers were more spontaneous compared to the second interview. The comparisons of alternatives in the second interview were considered as very unrealistic and extreme by SH7. However, the questions were more direct (SH4), which could be one reason why the interview was shorter (SH4, SH8). Thus, the methods used in the second interview seem to be better at helping people to make hard decisions faster (SH7, SH9).

5. Discussion

5.1. Design and application of the UTA$^{\mathrm{GMS}}$ *method*

One of our starting questions was, whether we could design and apply an elicitation procedure based on UTA$^{\mathrm{GMS}}$ for a very complex and large real case that allows for reasonably restricted interactions with busy stakeholders. To our knowledge, this is the first real-world application of UTA$^{\mathrm{GMS}}$ that demonstrates that it is indeed possible to have intensive, but not excessively time demanding interactions with a larger number of real stakeholders in a complex applied (environmental) decision problem, consisting of 19 objectives in this case. Our application has several highlights:

1) The procedure offered an interactive way to progressively elicit preferences. By computing the necessary preference relation from the stated preference, the unknown preferences for some pairs of alternatives were inferred. This substantially reduced the number of questions needed to rank the hypothetical alternatives (Tab. SM-5). For instance, we needed 16 hypothetical alternatives to elicit the preference for the main objective *Sustainable WWI*. After eliminating the weak dominance relations, 75 pairwise comparisons had to be asked to determine the complete ranking of the 16 alternatives. But on average we only used 9.8 questions (pairwise comparisons) thanks to the interactive inference of necessary preference relations. Moreover, we were able to choose the most informative question in each interaction, which also increased elicitation efficiency.

By chance, our heuristic is one of the heuristics studied in (Ciomek et al., 2017b), which was published after we had carried out our study and submitted this paper. We assumed that the probabilities of the stakeholders' answers for each pairwise question are equally likely. We maximized the estimated increase in the number of necessary preference relations. We only looked at the next elicitation method, i.e. used the search depth one, because using a greater search depth to increase accuracy would have involved significant additional costs. This simple heuristic appeared to perform sufficiently well in our case in terms of elicitation efficiency and computation time.

2) We elicited various forms of preference information. For each pairwise comparison, we requested the binary preference relation and the strength of the preference statement. The trade-off question, consisting of matching hypothetical alternatives with attributes at different levels, allowed us to generate an indifference statement. This enriched the preference information and was transformed to linear constraints when inferring MAVT models (section 3.3.4). By using different forms of questions, different aspects of the judgment process might be tapped (Huber et al., 1993).

3) Hypothetical reference alternatives were used, rather than the “real” alternatives of the case study for preference elicitation. The reason was that we wanted stakeholders to focus on the objectives, instead of on preferences about alternatives (value-focused thinking; Keeney, 1996).

However, the design of suitable reference alternatives requires more research. In this application, the constructed reference alternatives were very extreme and unrealistic, and the number of objectives exceeded two at several levels of the objectives hierarchy. This could partly explain why some stakeholders thought that the pairwise comparisons in the UTA$^{\text{GMS}}$ elicitation were very direct, but that the trade-offs between objectives seemed difficult. Some stakeholders stated that choosing between undesirable alternatives gave them very negative feelings compared to the SMART/SWING-variant, where they focused more strongly on improving objectives. Others have also observed that choosing from unattractive alternatives is difficult (Chatterjee and Heath, 1996; Schuwirth et al., 2012). Perhaps a different design of these reference alternatives would lead to a different perception of the pairwise comparisons. For example, Deparis et al. (2012) report that larger differences on each objective in the comparison of two alternatives increases the frequency of incomparability statements, when available. On the other hand, a larger magnitude of differences increases the use of indifference statements when only indifference and preference answers are permitted. Moreover, first studies indicate that the plausibility of hypothetical alternatives can have an effect on the consistency of the decision-makers responses; but there seem to be differences between intra- and inter-attribute preferences (van Valkenhoef and Tervonen, 2016; Vetschera et al., 2014). Thus, there is certainly opportunity for further research in this area.

5.2. Comparison of the two elicitation methods

Furthermore, we aimed at analyzing whether and in which way the two elicitation philosophies produce similar or diverging results in our case study application. We compared the two elicitation methods, direct elicitation of preference parameters and aggregation, or indirect inference of preferences by disaggregation from several points of view: the elicited MAVT model, based on preference statements (weights), the evaluation of alternatives (i.e. the outcome of the MCDA), and the perceptions of the stakeholders and analysts:

Weights. For individual stakeholders, the elicited weights differed between the two interviews, which has been observed before (Borcherding et al., 1991; Lienert et al., 2016; Pöyhönen and Hämäläinen, 2001). Even the ranking of the importance of main or sub-objectives sometimes changed between interviews. The observed variances in the weights were a consequence of mixed factors, which is discussed in more detail below (section 5.3).

Evaluation of alternatives.

The six alternatives were evaluated for both interviews and all ten stakeholders. We observed some disagreements between the performances of the alternatives, based on the stakeholders' preference statements. Moreover, there were some changes in the evaluation results between the first and second interview for the same stakeholder. For example, in the first interview (direct preference elicitation with SMART/SWING-variant), alternative A7 performed the best for stakeholders SH2–4, while being worst for stakeholder SH9. However, A7 became the second best for stakeholder SH9 in the second interview using UTA$^{\mathrm{GMS}}$ (Fig. 3). Statistically, the rankings of all six alternatives were significantly positively correlated for most stakeholders (Tab. 3). From the perspective of decision-making, the best-performing alternatives can be selected using the comprehensive values and rankings of these alternatives (Fig. 3). For example, we can count the number of stakeholders for whom the alternatives were ranked among the top three in both interviews. A8a is the best alternative, because it was among the top three alternatives for nine stakeholders in both interviews (Fig. 3). A7 and A8b followed by being ranked the top three for eight and five stakeholders, respectively, for both interviews. Alternatives A2, A5, and A9 seem to be inadequate and should be discarded (Fig. 3). Nevertheless, this selection is rather subjective. For example, if we were interested in the top two alternatives, A7 would be the best as it was among the top two alternatives for seven stakeholders compared to five stakeholders for A8a. We therefore recommend discussing A7 and A8a with the stakeholders as promising alternatives. For readers interested in a discussion of the wastewater infrastructure alternatives, we refer to our other publications in this case study (Lienert et al., 2016; Zheng et al., 2016).

The stakeholders. From the stakeholder feedback no final conclusion can be drawn concerning which elicitation method was easier, but it does provide some first insights. Although six stakeholders stated that the second interview (UTA$^{\text{GMS}}$) was easier, there were confounding factors for this (see below). However, some stakeholders gave feedback that the pairwise comparisons in the disaggregation elicitation (UTA$^{\text{GMS}}$) procedure were also cognitively demanding with difficult trade-offs. This somewhat contrasts claims that pairwise comparisons require less cognitive effort from the respondents (e.g. Greco et al., 2008; Kadziński et al., 2016; Kadziński and Tervonen, 2013; van Valkenhoef and Tervonen, 2016). We discussed in section 5.1 that the difficulty might be partly related to the design of hypothetical reference alternatives, which might seem extreme and unrealistic, and encourage more systematic research in this direction. However, it is worth pointing out that the perceived easiness is not necessarily the goal when designing an elicitation procedure. It has been suggested that the ease of judgment is bought at the cost of superficiality, and that easiness might preclude the potential benefits of thinking hard about the problem (Goodwin and Wright, 2001). Hoeffler and Ariely (1999) propose that more effort can lead to more reliable answers because it helps to construct preferences, which has received some support in an own related study (Lienert et al., 2016). Indeed, also in this study, several stakeholders explicitly said that the difficult questions in the UTA$^{\text{GMS}}$ interview helped them to make hard decisions (section 4.3).

We wish to point out two methodological drawbacks of our study design. First, for practical reasons our sample size of ten stakeholders is relatively small. It would not have been possible for us to carry out these intensive and time-demanding interviews with a larger number of stakeholders. Moreover, it was not possible to set up a fully designed experiment that allows for rigorous hypothesis testing, e.g. by controlling for influencing variables, in this complex real-world application. As stated in the Introduction, decision interventions in the real world require us to sacrifice on experimental hypothesis testing, making a generalization of the results difficult. However, we strongly believe, along with e.g. the recent literature from Behavioral Operational Research (BOR), that such more-qualitative but empirically grounded research provides valuable insights (e.g. Franco and Hämäläinen, 2016; Hämäläinen et al., 2013; Voinov et al., 2016). It can give us indications for future research (e.g. in more controlled settings) and provides important guidance on the limitations and potentials of MCDA for practitioners.

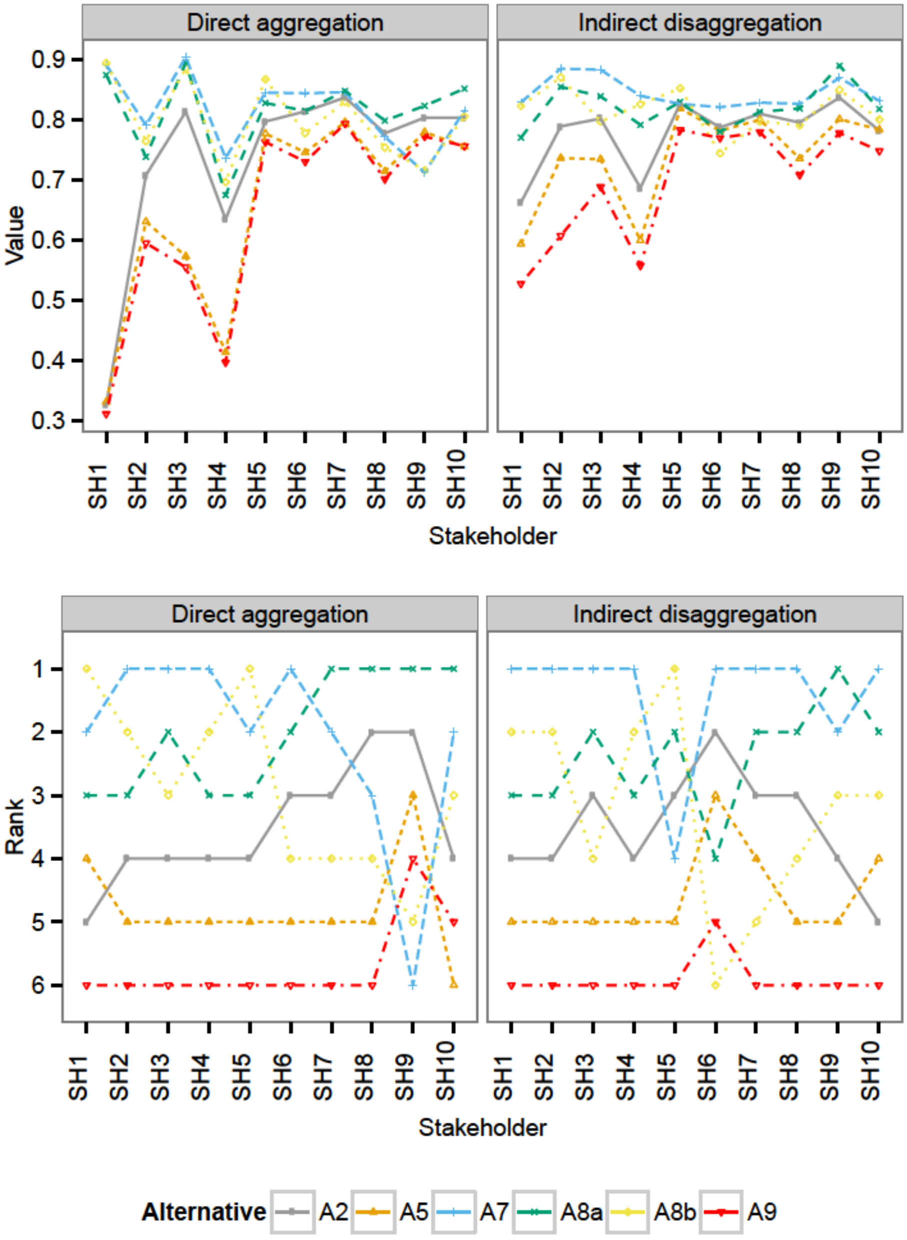

Figure 3: Values and ranks of six wastewater infrastructure alternatives (A2–A9) for ten stakeholders (SH1–SH10) and two elicitation methods: direct aggregation elicitation (SMART/SWING-variant for weights; bi-section method or rough elicitation for shape of marginal value functions), and indirect disaggregation elicitation (UTA$^{\mathrm{GMS}}$). y-axis of figure above: values of alternatives, where value 1 = all objectives are fully achieved and value 0= no objectives are achieved at all; y-axis of figure below: ranks of alternatives, where rank 1 = the alternative is the best among all six alternatives and rank 6 = the alternative is the worst among all six alternatives; x-axis: stakeholders (SH1–SH10).

Second, we consistently first carried out the interviews using the direct aggregation approach (i.e. elicitation of weights and value functions), followed by the indirect elicitation procedure (UTA$^{\text{GMS}}$) in the second interview. Thus, a confounding between increased familiarity with the problem, our types of questions concerning preferences, and with the stakeholder's perception about ease of use is possible. Indeed, several stakeholders confirmed that they had more insight into the decision problem in the second interview. We could have used a split sample design and randomized the order of interviews, but regarded a sample size of N=5 as too small to be able to draw better conclusions. We urge our colleagues to replicate such comparisons in other (real-world) applications, if possible using an improved experimental design based on larger sample sizes.

The analyst. From our own point of view, the UTA$^{\text{GMS}}$ interview clearly required more effort to prepare and implement the computer tool for the interactive elicitation procedure, compared to the preparation required for direct aggregation (i.e. the SMART/SWING-variant for weights and bi-section method or rough elicitation for shape of value functions). Moreover, the preference modeling and inference of weights and marginal value functions was more demanding for UTA$^{\text{GMS}}$ because of the linear programming problem. A future line of practice-oriented research might focus on designing suitable generalized UTA$^{\text{GMS}}$ software that can easily be adapted to different types of decision problems.

However, during elicitation, we found that the questions for pairwise comparisons were easier to explain than those for the SMART/SWING-variant to elicit weights and the bi-section method to elicit marginal value functions. The assessment of marginal value functions was particularly difficult. Furthermore, we noticed that people often ran into the goal-directed bias when weights were elicited with the SMART/SWING-variant, which has been observed by others (e.g. Schuwirth et al., 2012, when applying the "Reversed SWING" method). This should be noticed and corrected in an interview, but requires awareness and possibly some experience by the analyst.

One may wonder which kinds of methods should be chosen in which situation. We encourage combining different elicitation methods to avoid any systematic bias of a specific method (Hurson and Siskos, 2014; Montibeller and Winterfeldt, 2015) and to enhance the trustworthiness of the recommendation (Gregory et al., 2012; Hämäläinen et al., 2013; Lienert et al., 2016). According to Belton and Stewart (2002), elicitation methods inferring preferences from pairwise comparisons of alternatives (indirect methods in our terms) are more appropriate for conducting preliminary analysis, "quick and dirty" evaluation, and for making unimportant decisions. They suggest that returning to the direct analysis (direct aggregation methods in our terms) may well be necessary to reach a final conclusion. We are reluctant to agree with this statement, because some stakeholders gave the feedback that the pairwise comparisons of the UTA$^{\mathrm{GMS}}$ method required more effort, but also helped them to make the hard decisions required for this problem. We think that more research on the behavioral influences of the elicitation methods is necessary so that we better know how to integrate the two approaches.

5.3. Stability of preferences

Traditional economics assumes that people harbor well-articulated and stable preferences, and that preference elicitation consists of retrieving these pre-existing preferences. The constructive approach postulates that preferences are constructed based on the task and the context factors present during elicitation (e.g. Hoeffler and Ariely, 1999; Lichtenstein and Slovic, 2006; Slovic, 1995). The expressed preferences can thus depend on irrelevant context, such as task instructions, elicitation methods, information framing, and information order (Carlson and Bond, 2006). The economics literature uses test-retest experiments to study the reliability of e.g. contingent valuation surveys and choice experiments (e.g. Brouwer and Bateman, 2005; Liebe et al., 2012; Schaafsma et al., 2014). A fair to substantial reliability is generally found (Liebe et al., 2012). But the research on preference stability regarding MCDA is scarce. In an own related study, we used a multi-method elicitation approach with stakeholders and the general public, and we repeated elicitation after one month (Lienert et al., 2016). We found that the weight elicitation method was the most important predictor of preference stability over time.

This application in the paper presented here provides an empirical and qualitative study on preference stability for MCDA through in-depth interviews. The observed inconsistency of preferences of individual stakeholders across the two interviews along with the explanations for their inconsistency provides strong evidence that the preferences were constructed rather than pre-existing. Hereby, three factors seem especially relevant:

1) The preferences of the stakeholders were influenced by external factors. Our ten stakeholders were identified by a rigid stakeholder analysis, combined with a social network analysis (Lienert et al., 2013) and at least some are very experienced in the wastewater domain. Despite this, in some cases external influences likely caused changes in their underlying preferences. For example, the sub-objective *Fewer structural failures of drainage system* was more important to stakeholder SH1 in the second interview, as

there was a pipe failure in his neighborhood between the two interviews (see Section 4.3.1).

2) We observed a learning effect especially for those stakeholders who were uncertain about their preferences. There were frequent shifts from an indifference statement in the first interview to a preference statement in the second (Tab. 2). Besides being a method effect, we suspect that this is also related to learning. Indeed, the stakeholders frequently admitted that their preference change could stem from their uncertainty about their own preferences and that there had been a learning effect in between the interviews. This corresponds to the tendency of people to avoid making trade-offs, which is a common response to cognitively and emotionally demanding tasks (Payne et al., 1999). Possibly, people were more reluctant to make difficult trade-offs in the first interview, because they were not yet used to the decision context and these types of questions.

3) Irrelevant context can also explain many cases of preference instability between the two interviews. We identified two of the contextual factors in our case study application. The first is the elicitation method and the second the different decision strategies which the respondents used when answering questions. We have provided evidence that the stakeholders perceived the two elicitation methods differently and might have changed their preferences due to different questions in the interview (Section 4.3.2). However, it is difficult to know precisely why they used different strategies, as preferences can be influenced by various irrelevant factors, such as option "framing", changes in the "choice context", or the presence of prior cues or "anchors" (Lichtenstein and Slovic, 2006). We wish to point out that a different elicitation method might also evoke a different decision strategy. For example, stakeholder SH5 paid more attention to the negative consequences on human beings in the second interview and therefore changed his preference. This could be because the pairwise comparisons of UTA^{GMS} asking the respondents to choose between undesirable alternatives give a negative feeling. Thus, our categorization of these two factors in Table SM-5.2 is to some extent subjective and the categories can also overlap.

The constructive approach of preference formation indicates that we should pay more attention to facilitate preference construction and learning in the elicitation process. However, this is usually not sufficiently emphasized in practical decision-making (see Karjalainen et al., 2013, for an example). We believe that learning in MCDA processes might be more important for their success than simply obtaining a ranking of alternatives, as stated by others (Karjalainen et al., 2013; Marttunen et al., 2015).

6. Conclusions

The aim of this paper was to register the implementation impacts of two MCDA preference elicitation philosophies (aggregation and disaggregation) in a large, real-world case and give guidance on these elicitation approaches for practitioners. The elicitation methods consist of a typically used direct aggregation elicitation procedure versus an indirect disaggregation elicitation procedure. We carried out two sets of interviews with ten real stakeholders, who were willing to give in-depth preference statements twice, purely for the sake of research. In the interview using direct aggregation methods, a SMART/SWING-variant was used for weight elicitation, and a combination of the bi-section method and rough elicitation of the shape to elicit the marginal value functions. For indirect disaggregation, an elicitation procedure based on pairwise comparisons of hypothetical alternatives using the UTA$^{\mathrm{GMS}}$ method was implemented to simultaneously assess the weights and marginal value functions from indirect preference information. To our knowledge, such large real environmental applications of the relatively new UTA$^{\mathrm{GMS}}$ are scarce or inexistent. By analyzing the results of the two interviews, we found that:

1) The indirect aggregation elicitation procedure based on the UTA$^{\mathrm{GMS}}$ method is applicable for complex real-world cases. We encourage more applications to more rigorously test this finding. We believe that the use of indirect disaggregation methods will greatly enrich the practice of MCDA.

2) The two elicitation procedures were perceived differently by both the respondents and the analysts. For the stakeholders, the pairwise comparisons with difficult trade-offs still seemed cognitively demanding, which somewhat contradicts earlier statements of some researchers (e.g. Greco et al., 2008; Kadziński et al., 2016; Kadziński and Tervonen, 2013; van Valkenhoef and Tervonen, 2016). Additionally, we have indications that such difficult questions might even help the stakeholders to better form their preferences (Hoeffler and Ariely, 1999). For the analyst, the UTA$^{\mathrm{GMS}}$ procedure required more effort to develop the supporting computer tool for interactive elicitation and some knowledge about linear programming is necessary for the inference of the MAVT models. On the other hand, the pairwise comparisons in the UTA$^{\mathrm{GMS}}$ method were easy to explain to the stakeholders, while the bi-section method to elicit marginal value functions was especially difficult to use. More systematic investigation with larger sample sizes and in other application contexts is needed to further compare the approaches.

3) We observed that preferences were evidently constructed during the elicitation procedure. This indicates the necessity of taking into account the

constructive nature of preferences in the application of MCDA to complex environmental decisions, which is, however, missing in many applications. Supporting stakeholders to construct preferences and enhancing learning during the decision-making process might be more important than using complicated decision models to reach a "correct" conclusion.

7. Acknowledgments

This research has been carried out within the National Research Program "Sustainable Water Management" (NRP 61; www.nfp61.ch; project number: 406140_125901 / 1), funded by the Swiss National Science Foundation. We would like to thank Christoph Egger, Max Maurer, Peter Reichert, Lisa Scholten, Nele Schuwirth, and Anja Zorn for their valuable contributions. Special thanks to the stakeholders for their participation in the case study and their willingness to carry out a second interview purely for research purposes.

Supplementary material

Supplementary material associated with this article can be found, in the online version at XXX.

Appendix A. Lowest level sub-objectives

Table A.1: Definitions of the lowest level sub-objectives (Fig. 1).

Shortname	Sub-objectives
rehab	Low future rehabilitation burden until 2050
adapt	Flexible system adaptation
chem	Good chemical state of watercourses
hydr	Low negative hydraulic impacts
exfiltrsew	Low contamination from sewers exfiltration
exfiltrstruct	Low contamination from infiltration structures
phosph	Recovery of nutrients
econs	Efficient use of electrical energy
illn	Few gastrointestinal infections through direct contact with wastewater (failures of infrastructures)
cso	Few gastro-intestinal infections through indirect contact with wastewater (swimming after CSOs)
failure	Few structural failures of drainage system
service	Sufficient drainage capacity of drainage system
efqm	High quality of management and operations
voice	High co-determination of citizens in infrastructure decisions
time	Low time demand for end user
area	Low additional area demand for end user
collab	Low unnecessary construction and road works
costcap	Low annual costs
costchange	Low cost increase

References

Amerise, I.L., Marozzi, M., Tarsitano, A., 2015. pvrank: Rank Correlations. R package version 1.0.

Anderson, R.M., Clemen, R., 2013. Toward an improved methodology to construct and reconcile decision analytic preference judgments. Decision Analysis 10, 121–134. doi:10.1287/deca.2013.0268.

Bana, C.A., De Corte, J.M., Vansnick, J.C., 2016. On the mathematical foundations of MACBETH, in: Multiple Criteria Decision Analysis. Springer, pp. 421–463. doi:10.1007/0-387-23081-5_10.

Behzadian, M., Kazemzadeh, R., Albadvi, A., Aghdasi, M., 2010. PROMETHEE: A comprehensive literature review on methodologies and applications. European Journal of Operational Research 200, 198–215. doi:10.1016/j.ejor.2009.01.021.

Beinat, E., 1997. Value functions for environmental management. Springer.

Belton, V., 1986. A comparison of the analytic hierarchy process and a simple multi-attribute value function. European Journal of Operational Research 26, 7–21. doi:10.1016/0377-2217(86)90155-4.

Belton, V., Stewart, T., 2002. Multiple Criteria Decision Analysis: An Integrated Approach. Kluwer.

Beuthe, M., Scannella, G., 2001. Comparative analysis of UTA multicriteria methods. European Journal of Operational Research 130, 246–262. doi:10.1016/S0377-2217(00)00042-4.

Borcherding, K., Eppel, T., von Winterfeldt, D., 1991. Comparison of weighting judgments in multiattribute utility measurement. Management Science 37, 1603–1619. doi:10.1287/mnsc.37.12.1603.

Bous, G., Fortemps, P., Glineur, F., Pirlot, M., 2010. ACUTA: A novel method for eliciting additive value functions on the basis of holistic preference statements. European Journal of Operational Research 206, 435–444. doi:10.1016/j.ejor.2010.03.009.

Branke, J., Corrente, S., Greco, S., Gutjahr, W., 2017. Efficient pairwise preference elicitation allowing for indifference. Computers & Operations Research 88, 175–186. doi:10.1016/j.cor.2017.06.020.

Brans, J., Vincke, P., Mareschal, B., 1986. How to select and how to rank projects: The PROMETHEE method. European Journal of Operational Research 24, 228–238. doi:10.1016/0377-2217(86)90044-5.

Brouwer, R., Bateman, I.J., 2005. Temporal stability and transferability of models of willingness to pay for flood control and wetland conservation. Water Resources Research 41, 1–6. doi:10.1029/2004WR003466.

Carlson, K.A., Bond, S.D., 2006. Improving preference assessment: limiting the effect of context through pre-exposure to attribute levels. Management Science 52, 410–421. doi:10.1287/mnsc.1050.0434.

Chatterjee, S., Heath, T.B., 1996. Conflict and loss aversion in multiattribute choice: the effects of trade-off size and reference dependence on decision difficulty. Organizational Behavior & Human Decision Processes 67, 144–155. doi:10.1006/obhd.1996.0070.

Cinelli, M., Coles, S.R., Kirwan, K., 2014. Analysis of the potentials of multi criteria decision analysis methods to conduct sustainability assessment. Ecological Indicators 46, 138–148. doi:10.1016/j.ecolind.2014.06.011.

Ciomek, K., Kadziński, M., Tervonen, T., 2017a. Heuristics for prioritizing pair-wise elicitation questions with additive multi-attribute value models. Omega 71, 27–45. doi:10.1016/j.omega.2016.08.012.

Ciomek, K., Kadziński, M.K.M., Tervonen, T., 2017b. Heuristics for selecting pair-wise elicitation questions in multiple criteria choice problems. European Journal of Operational Research 262, 693–707. doi:10.1016/j.ejor.2017.04.021.

Corrente, S., Greco, S., Slowinski, R., 2012. Multiple Criteria Hierarchy Process in Robust Ordinal Regression. Decision Support Systems 53, 660–674. doi:10.1016/j.dss.2012.03.004.

Deparis, S., Mousseau, V., Ozturk, M., Pallier, C., Huron, C., 2012. When conflict induces the expression of incomplete preferences. European Journal of Operational Research 221, 593–602. doi:10.1016/j.ejor.2012.03.041.

Diakoulaki, D., Zopounidis, C., Mavrotas, G., Doumpos, M., 1999. The use of a preference disaggregation method in energy analysis and policy making. Energy 24, 157–166. doi:10.1016/S0360-5442(98)00081-4.

Doumpos, M., Xidonas, P., Xidonas, S., Siskos, Y., 2016. Development of a robust multicriteria classification model for monitoring the postoperative behaviour of heart patients. Journal of Multi-Criteria Decision Analysis 23, 15–27. doi:10.1002/mcda.1547.

Doumpos, M., Zanakis, S.H., Zopounidis, C., 2001. Multicriteria preference disaggregation for classification problems with an application to global investing risk. Decision Sciences 32, 333–386. doi:10.1111/j.1540-5915.2001.tb00963.x.

Doumpos, M., Zopounidis, C., 2011. Preference disaggregation and statistical learning for multicriteria decision support: A review. European Journal of Operational Research 209, 203–214. doi:10.1016/j.ejor.2010.05.029.

Eisenführ, F., Martin, W., Thomas, L., 2010. Rational Decision Making. 1st ed., Springer Verlag, Berlin, Heidelberg, New York.

Ferretti, V., Bottero, M., Mondini, G., 2014. Decision making and cultural heritage: An application of the Multi-Attribute Value Theory for the reuse of historical buildings. Journal of Cultural Heritage 15, 644–655. doi:10.1016/j.culher.2013.12.007.

Figueira, J., Greco, S., Ehrogott, M. (Eds.), 2016. Multiple Criteria Decision Analysis: State of the Art Surveys. volume 233. 2 ed., Springer-Verlag New York.

Figueira, J.R., Greco, S., Roy, B., Słowiński, R., 2013. An overview of ELECTRE methods and their recent extensions. Journal of Multi-Criteria Decision Analysis 20, 61–85. doi:10.1002/mcda.1482.

Figueira, J.R., Greco, S., Slowiński, R., 2009. Building a set of additive value functions representing a reference preorder and intensities of preference: GRIP method. European Journal of Operational Research 195, 460–486. doi:10.1016/j.ejor.2008.02.006.

Franco, L.A., Hämäläinen, R.P., 2016. Behavioural operational research: Returning to the roots of the OR profession. European Journal of Operational Research 249, 791–795. doi:10.1016/j.ejor.2015.10.034.

Ghaderi, M., Ruiz, F., Agell, N., 2015. Understanding the impact of brand colour on brand image: A preference disaggregation approach. Pattern Recognition Letters 67, 11–18. doi:10.1016/j.patrec.2015.05.011.

Gleick, P.H., 2003. Global freshwater resources: soft-path solutions for the 21st century. Science 302, 1524–1528. doi:10.1126/science.1089967.

Goodwin, P., Wright, G., 2001. Enhancing strategy evaluation in scenario planning: a role for decision analysis. Journal of Management Studies 38, 1–16. doi:10.1111/1467-6486.00225.

Greco, S., Matarazzo, B., Slowiński, R., 2001. Rough sets theory for multicriteria decision analysis. European Journal of Operational Research 129, 1–47. doi:10.1016/S0377-2217(00)00167-3.

Greco, S., Mousseau, V., Slowiński, R., 2008. Ordinal regression revisited: multiple criteria ranking using a set of additive value functions. European Journal of Operational Research 191, 416–436. doi:10.1016/j.ejor.2007.08.013.

Greco, S., Mousseau, V., Slowinski, R., 2014. Robust ordinal regression for value functions handling interacting criteria. European Journal of Operational Research 239, 711–730. doi:10.1016/j.ejor.2014.05.022.

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.

Hajkowicz, S., Collins, K., 2007. A review of multiple criteria analysis for water resource planning and management. Water resources management 21, 1553–1566. doi:10.1007/s11269-006-9112-5.

Hoeffler, S., Ariely, D., 1999. Constructing stable preferences: A look into dimensions of experience and their impact on preference stability. Journal of Consumer Psychology 8, 113–139. doi:10.1207/s15327663jcp0802_01.

Huang, I.B., Keisler, J., Linkov, I., 2011. Multi-criteria decision analysis in environmental sciences: Ten years of applications and trends. Science of The Total Environment 409, 3578–3594. doi:10.1016/j.scitotenv.2011.06.022.

Huber, J., Wittink, D.R., Fiedler, J.A., Miller, R., 1993. The effectiveness of alternative preference elicitation procedures in predicting choice. Journal of Marketing Research 30, 105–114. doi:10.2307/3172685.

Hurson, C., Siskos, Y., 2014. A synergy of multicriteria techniques to assess additive value models. European Journal of Operational Research 238, 540–551. doi:10.1016/j.ejor.2014.03.047.

Hämäläinen, R.P., 2015. Behavioural issues in environmental modelling - The missing perspective. Environmental Modelling & Software 73, 244–253. doi:10.1016/j.envsoft.2015.08.019.

Hämäläinen, R.P., Luoma, J., Saarinen, E., 2013. On the importance of behavioral operational research: The case of understanding and communicating about dynamic systems. European Journal of Operational Research 228, 623–634. doi:10.1016/j.ejor.2013.02.001.

Iyengar, V.S., Lee, J., Campbell, M., 2001. Evaluating multiple attribute items using queries, in: Proceedings of the 3rd ACM Conference on Electronic Commerce, ACM, New York, NY, USA. pp. 144–153. doi:10.1145/501158.501174.

Jacquet-Lagreze, E., Siskos, J., 1982. Assessing a set of additive utility functions for multicriteria decision-making, the UTA method. European Journal of Operational Research 10, 151–164. doi:10.1016/0377-2217(82)90155-2.

Jacquet-Lagreze, E., Siskos, Y., 2001. Preference disaggregation: 20 years of MCDA experience. European Journal of Operational Research 130, 233–245. doi:10.1016/S0377-2217(00)00035-7.

Kadziński, M., Cinelli, M., Ciomek, K., Coles, S.R., Nadagouda, M.N., Varma, R.S., Kirwan, K., 2016. Co-constructive development of a green chemistry-based model for the assessment of nanoparticles synthesis. European Journal of Operational Research doi:10.1016/j.ejor.2016.10.019.

Kadziński, M., Tervonen, T., 2013. Robust multi-criteria ranking with additive value models and holistic pair-wise preference statements. European Journal of Operational Research 228, 169–180. doi:10.1016/j.ejor.2013.01.022.

Kadziński, M., Ciomek, K., Rychły, P., Słowiński, R., 2016. Post factum analysis for robust multiple criteria ranking and sorting. Journal of Global Optimization 65, 531–562. doi:10.1007/s10898-015-0359-3.

Karjalainen, T., Rossi, P., Ala-aho, P., Eskelinen, R., Klove, B., Pulido-Velazquez, M., Yang, H., 2013. A decision analysis framework for stakeholder involvement and learning in groundwater management. Hydrology and Earth System Sciences 17, 1–13. doi:10.5194/hess-17-5141-2013.

Keeney, R.L., 1996. Value-focused thinking: Identifying decision opportunities and creating alternatives. European Journal of Operational Research 92, 537–549. doi:10.1016/0377-2217(96)00004-5.

Keeney, R.L., Raiffa, H., 1976. Decisions with multiple objectives : preferences and value tradeoffs. Wiley.

Kendall, M.G., 1938. A new measure of rank correlation. Biometrika 30, 81–93. doi:10.2307/2332226.

Känel, B., Niederhauser, P., Meier, W., 2008. “Zustand der Fliessgewässer in den Einzugsgebieten von Sihl, Limmat und Zürichsee; Messkampagne 2006 / 2007”; in German (State of watercourses in the catchments of Sihl, Limmatt, and Lake Zürich; measurement campaign 2006 / 2007).

Technical Report. AWEL, Amt für Abfall, Wasser, Energie und Luft, Kt. Zürich (Office for waste, water, energy, and air, ct. Zürich). URL: http://www.gewaesserschutz.zh.ch/. Accessed 29. June 2017.

Langhans, S.D., Lienert, J., 2016. Four common simplifications of multi-criteria decision analysis do not hold for river rehabilitation. Plos One 11. doi:10.1371/journal.pone.0150695.

Langhans, S.D., Reichert, P., Schuwirth, N., 2014. The method matters: a guide for indicator aggregation in ecological assessments. Ecological Indicators 45, 494–507. doi:10.1016/j.ecolind.2014.05.014.

Larsen, T.A., Hoffmann, S., Lüthi, C., Truffer, B., Maurer, M., 2016. Emerging solutions to the water challenges of an urbanizing world. Science 352, 928–933. doi:10.1126/science.aad8641.

Lichtenstein, S., Slovic, P., 2006. The construction of preference. Cambridge University Press, New York.

Liebe, U., Meyerhoff, J., Hartje, V., 2012. Test-retest reliability of choice experiments in environmental valuation. Environmental and Resource Economics 53, 389–407. doi:10.1007/s10640-012-9567-1.

Lienert, J., Duygan, M., Zheng, J., 2016. Preference stability over time with multiple elicitation methods to support wastewater infrastructure decision-making. European Journal of Operational Research 253, 746–760. doi:10.1016/j.ejor.2016.03.010.

Lienert, J., Koller, M., Konrad, J., McArdell, C.S., Schuwirth, N., 2011. Multiple-criteria decision analysis reveals high stakeholder preference to remove pharmaceuticals from hospital wastewater. Environmental Science & Technology 45, 3848–3857. doi:10.1021/es1031294.

Lienert, J., Schnetzer, F., Ingold, K., 2013. Stakeholder analysis combined with social network analysis provides fine-grained insights into water infrastructure planning processes. Journal of Environmental Management 125, 134–48. doi:10.1016/j.jenvman.2013.03.052.

Lienert, J., Scholten, L., Egger, C., Maurer, M., 2015. Structured decision-making for sustainable water infrastructure planning and four future scenarios. EURO Journal on Decision Processes 3, 107–140. doi:10.1007/s40070-014-0030-0.

Macharis, C., Springael, J., Brucker, K.D., Verbeke, A., 2004. PROMETHEE and AHP: The design of operational synergies in multicriteria analysis. European Journal of Operational Research 153, 307–317. doi:10.1016/S0377-2217(03)00153-X.

Marttunen, M., Belton, V., Lienert, J., 2017. Are objectives hierarchy related biases observed in practice? A meta-analysis of environmental and energy applications of multi-criteria decision analysis. European Journal of Operational Research doi:10.1016/j.ejor.2017.02.038.

Marttunen, M., Mustajoki, J., Dufva, M., Karjalainen, T.P., 2015. How to design and realize participation of stakeholders in MCDA processes? A framework for selecting an appropriate approach. Euro Journal on Decision Processes 3, 187–214. doi:10.1007/s40070-013-0016-3.

Mendoza, G., Martins, H., 2006. Multi-criteria decision analysis in natural resource management: A critical review of methods and new modelling paradigms. Forest Ecology and Management 230, 1–22. doi:10.1016/j.foreco.2006.03.023.

Milly, P.C.D., Betancourt, J., Falkenmark, M., Hirsch, R.M., Kundzewicz, Z.W., Lettenmaier, D.P., Stouffer, R.J., 2008. Stationarity is dead: whither water management? Science 319, 573–574. doi:10.1126/science.1151915.

Montibeller, G., Winterfeldt, D.V., 2015. Cognitive and motivational biases in decision and risk analysis. Risk Analysis 35, 1230–1251. doi:10.1111/risa.12360.

Mustajoki, J., Hämäläinen, R.P., Salo, A., 2005. Decision support by interval SMART/SWING-incorporating imprecision in the SMART and SWING methods. Decision Sciences 36, 317–339. doi:10.1111/j.1540-5414.2005.00075.x.

Mustajoki, J., Saarikoski, H., Marttunen, M., Ahtikoski, A., Hallikainen, V., Helle, T., Hyppönen, M., Jokinen, M., Naskali, A., Tuulentie, S., Varmola, M., Vatanen, E., Ylisirniö, A.L., 2011. Use of decision analysis interviews to support the sustainable use of the forests in Finnish Upper Lapland. Journal of Environmental Management 92, 1550–1563. doi:10.1016/j.jenvman.2011.01.007.

Payne, J., Bettman, J., Schkade, D., 1999. Measuring Constructed Preferences: Towards a Building Code. Journal of Risk and Uncertainty 19, 243–270. doi:10.1023/A:1007843931054.

Payne, J.W., Bettman, J.R., Johnson, E.J., 1992. Behavioral decision research: A constructive processing perspective. Annual review of psychology 43, 87–131. doi:10.1146/annurev.ps.43.020192.000511.

Pöyhönen, M., Hämäläinen, R.P., 2001. On the convergence of multiattribute weighting methods. European Journal of Operational Research 129, 569–585. doi:10.1016/S0377-2217(99)00467-1.

R Development Core Team, 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org.

Reichert, P., Langhans, S.D., Lienert, J., Schuwirth, N., 2015. The conceptual foundation of environmental decision support. Journal of Environmental Management 154, 316–332. doi:10.1016/j.jenvman.2015.01.053.

Reichert, P., Schuwirth, N., Langhans, S., 2013. Constructing, evaluating and visualizing value and utility functions for decision support. Environmental Modelling & Software 46, 283–291. doi:10.1016/j.envsoft.2013.01.017.

Roy, B., 1996. Multicriteria Methodology for Decision Aiding. Kluwer Academic, Dordrecht.

Roy, B., Bouyssou, D., 1993. Aide Multicritère à la Décision: Méthodes et Cas. Economica, Paris.

Saaty, T., 1980. The Analytic Hierarchy Process. McGraw-Hill.

Schaafsma, M., Brouwer, R., Liekens, I., de Nocker, L., 2014. Temporal stability of preferences and willingness to pay for natural areas in choice experiments: a test-retest. Resource and Energy Economics 38, 243–260. doi:10.1016/j.reseneeco.2014.09.001.

Scholten, L., Schuwirth, N., Reichert, P., Lienert, J., 2015. Tackling uncertainty in multi-criteria decision analysis: An application to water supply infrastructure planning. European Journal of Operational Research 242, 243–260. doi:10.1016/j.ejor.2014.09.044.

Schuwirth, N., Reichert, P., Lienert, J., 2012. Methodological aspects of multi-criteria decision analysis for policy support: A case study on pharmaceutical removal from hospital wastewater. European Journal of Operational Research 220, 472–483. doi:10.1016/j.ejor.2012.01.055.

Siskos, Y., Grigoroudis, E., Matsatsinis, N.F., 2016. UTA Methods, in: Greco, S., Ehrgott, M., Figueira, J.R. (Eds.), Multiple Criteria Decision Analysis: State of the Art Surveys. Springer, New York. volume 233, pp. 315–362. doi:10.1007/978-1-4939-3094-4_9.

Slovic, P., 1995. The construction of preference. American psychologist 50, 364. doi:10.1037//0003-066x.50.5.364.

Smith, J.E., von Winterfeldt, D., 2004. Anniversary article: Decision analysis in management science. Management Science 50, 561–574. doi:10.1287/mnsc.1040.0243.

Spyridakos, A., Siskos, Y., Yannacopoulos, D., Skouris, A., 2001. Multicriteria job evaluation for large organizations. European Journal of Operational Research 130, 375–387. doi:10.1016/S0377-2217(00)00039-4.

Tervonen, T., 2013. ror: Robust Ordinal Regression MCDA library. URL: http://CRAN.R-project.org/package=ror.

Toubia, O., Hauser, J.R., Simester, D.I., 2004. Polyhedral methods for adaptive choice-based conjoint analysis. Journal of Marketing Research 41, 116–131. doi:10.1509/jmkr.41.1.116.25082.

van Valkenhoef, G., Tervonen, T., 2016. Entropy-optimal weight constraint elicitation with additive multi-attribute utility models. Omega 64, 1–12. doi:10.1016/j.omega.2015.10.014.

Vetschera, R., Weitzl, W., Wolfsteiner, E., 2014. Implausible alternatives in eliciting multi-attribute value functions. European Journal of Operational Research 234, 221–230. doi:10.1016/j.ejor.2013.09.016.

Voinov, A., Kolagani, N., McCall, M.K., Glynn, P.D., Kragt, M.E., Ostermann, F.O., Pierce, S.A., Ramu, P., 2016. Modelling with stakeholders – Next generation. Environmental Modelling & Software 77, 196–220. doi:10.1016/j.envsoft.2015.11.016.

Weber, M., Borcherding, K., 1993. Behavioral influences on weight judgments in multiattribute decision making. European Journal of Operational Research 67, 1–12. doi:10.1016/0377-2217(93)90318-H.

Zheng, J., Egger, C., Lienert, J., 2016. A scenario-based MCDA framework for wastewater infrastructure planning under uncertainty. Journal of Environmental Management 183, 895–908. doi:10.1016/j.jenvman.2016.09.027.

Zopounidis, C., 2001. Preference disaggregation in financial modeling: Basic features and some examples. Operational Research 1, 263–284. doi:10.1007/BF02936355.

Zopounidis, C., Doumpos, M., Zanakis, S., 2007. Stock evaluation using a preference disaggregation methodology. Decision Sciences 30, 313–336. doi:10.1111/j.1540-5915.1999.tb01612.x.