Gamified online survey to elicit citizens' preferences and enhance learning for environmental decisions

Alice H. Aubert ^{a*}; Judit Lienert ^a

^a Eawag: Swiss Federal Institute of Aquatic Science and Technology, Überlandstrasse 133, CH-8600 Dübendorf, Switzerland

* Corresponding author: Telephone +41(0)58 765 56 88, Fax: +41(0)58 765 58 02, e-mail: alice.aubert@eawag.ch

Highlights

- Proof of concept of a gamified online weight elicitation survey for MCDA
- Two-level learning loops enable factual learning and preference construction
- Survey positively perceived, particularly with story and non-player characters
- Suggestions to improve engagement in the gamified survey are discussed
- Survey to enhance citizen participation in complex environmental decision-making

Abstract

Multi-Criteria Decision Analysis (MCDA) requires a critical step, namely to elicit individual preferences. On the basis of learning theories, we formalize preference construction as learning about facts and values, and as a process; we also conceptualize an online preference elicitation survey that offers learning loops to increase factual learning and support preference construction. Another originality is gamification. Game elements (a narrative and non-player characters as motivational affordance) keep respondents engaged in the demanding task of weight elicitation. Our tool enables broad public participation in MCDA, allowing reliable online preference elicitation. The survey concept was tested with 107 students and a control treatment. Quantitative and qualitative data indicate that the concept works. Participants' factual knowledge increased. The survey helped students to learn about their own preferences concerning the importance of objectives. The practical implication is that weighting can be reliably elicited by online surveys. Participants reported a positive experience; further ways to improve it are thoroughly discussed.

Gamified online survey to elicit citizens' preferences and enhance learning for environmental decisions

eawag
aquatic research

When answering the survey:

- Participants learn *facts about wastewater*
- Participants learn about *their own preferences*
- Participants have a *positive experience*

Keywords

Environmental decision support systems, gamification, learning, preference construction, online survey, participatory decision-making

1. Introduction

1.1. Research question and objectives

This paper focuses on preference elicitation for major, long-term, publicly funded environmental decision-making. Such decisions may significantly impact the three pillars of sustainability as well as citizens' daily life. These decisions ought to be more participatory than at present, involving the affected citizens (Aubert et al., 2018). As decision analysts, we aim to elicit "reflective responses in an active process of preference construction" (Fischhoff, 2006). To date, this is mostly done during face-to-face interviews or group workshops, facilitated by the decision analyst. His/her role is to (1) ensure that the problem is understood, (2) structure the process, (3) collect the participants' values given the specific decision context, and (4) translate the collected preferences into a useful format (adapted from Fischhoff, 2006). Attempts to make preference elicitation for Multi-Criteria Decision Analysis (MCDA) more participatory using online technology have been made (e.g. Hämäläinen et al., 2010; Lienert et al., 2016; Mustajoki et al., 2004). However, they are still rare and critically received: without the guidance of a facilitator, participants may misunderstand the decision context and the complex tasks related to preference elicitation (Marttunen and Hämäläinen, 2008). The well-established procedures for MCDA preference elicitation are cognitively demanding and repetitive (Eisenführ et al., 2010; Riabacke et al., 2012); this can enhance the occurrence of unwanted biases (Montibeller and von Winterfeldt, 2015).

We consequently developed a new online preference elicitation survey concept. It is meant to (1) provide the necessary information to understand the decision context so that participants can learn about the topics at stake, (2) enhance reflective thinking so that participants can construct their preferences, and (3) provide a positive self-motivating experience so that the difficulty and repetitiveness of the task are overcome. The survey concept is summarized in Section 2. This paper presents an experimental evaluation of a prototype of the survey concept. It aims to answer the overarching research question, i.e. is it possible to elicit many reliable preferences using online technology? In particular, does this prototype for weight elicitation meet the three goals for which it was designed?

The prototype of the survey concept has been tested for the three hypotheses described above. Results show that participants learn about the decision topic through the survey, that

the reflective approach during decision-making is enhanced, and that the experience of answering the survey is positively perceived overall. The causality with respect to the construction of preferences and the reliability of the elicited preferences was not proven, but is discussed in Section 5.

1.2. Summary of key related research

1.2.1. Participatory environmental decisions

Since the early calls in the 1990s (Fiorino, 1990; Renn et al., 1995), there is a growing body of policy-making literature in general, and specifically of environmental literature, arguing for citizen participation (e.g. Dupuis and Knoepfel, 2015; Papadopoulos and Warin, 2007, and references herein). Citizen participation, also referred to as including “ordinary citizens”/ “lay people”/ the “profane” (Papadopoulos and Warin, 2007) in decision-making, is expected to break gridlocks and avoid litigation following the decision. Participation should enhance a two-way education process by providing opportunities (1) for citizens to learn about the complex trade-offs involved in the decisions, and (2) for decision-makers to learn about “explosively unpopular” outcomes and to anticipate or avoid them (Irvin and Stansbury, 2004).

However, involving citizens in decision-making can be costly in terms of money and time. If the role of the citizens is not clearly defined or the method not suitably adapted, it can even worsen hostility towards the decision and decision-makers (Irvin and Stansbury, 2004). Time is particularly critical when citizens engage in deliberative democracy via focus groups, policy fora, or deliberative arenas, which entail real meetings (as opposed to virtual ones). In these personal deliberative approaches, intended to enhance the construction of individual preferences and reduce bounded rationality, a few tens of representative citizens are usually invited. This raises the principal-agent issue, i.e. how to achieve equal representation of the various interests. Thus, some authors favor a participatory democracy approach, meant to raise public participation by enabling a larger group of people to vote, but dedicating less time to the public debate (Papadopoulos and Warin, 2007). Clearly, deliberative and participatory approaches do not conflict and would benefit from each other.

The decision pathway survey (Gregory et al., 1997; Gregory et al., 2016) appears to be an attempt to bridge the deliberative and participatory approaches by (1) providing factual topical information to “arrive at a more complete understanding of [the citizens’] own values” (Gregory et al., 2016), and (2) involving a large sample of stakeholders thanks to web-based technology. Additionally, a decision pathway survey considers the behavioral issues

associated with decision-making by persons unfamiliar with the topic. It builds on the theory of the construction of preference (Lichtenstein and Slovic, 2006; Slovic, 1995).

1.2.2. The construction of preference

The theories of preference construction initially stem from behavioral decision research. They were mostly developed for choices in the context of consumer behaviors (Lichtenstein and Slovic, 2006). The various theories share the idea that preferences are constructed (Slovic, 1995) given the context, as opposed to pre-existing and retrieved, as had been assumed by classical economists. According to this theory, preference construction is context-dependent and can also be influenced by the principal reason for making the decision, by the cognitive capabilities of the participants and by their experience (Amir and Levav, 2008; Warren et al., 2011). Memory and learning, i.e. retrieving old information from memory and encoding new information in memory, at both cognitive and affective levels, contribute to preference construction (for a review, see Weber and Johnson, 2009). The way the problem is framed and the way the preference elicitation task is developed (wording, visualization, cognitive demand) influences the preference construction (Hoeffler and Ariely, 1999; Jorgensen et al., 2004; Liebe et al., 2012; Payne et al., 1992).

Decision-making on complex problems, e.g. decisions involving common environmental resources or any public resource, is challenged to take preferences from individuals into account. As a decision becomes too complex, individuals tend to simplify it and construct preferences using only a subset of all the available information: they use heuristics (Gregory et al., 2006). In practice, when using MCDA methods (e.g. Multi Attribute Value/ Utility Theory), complex problems are decomposed into smaller, manageable tasks (Keeney and Raiffa, 1976). The MCDA process is composed of (1) problem structuring, (2) assessing facts (predictions, a_i) and values (preferences, $v_i(a_i)$ and w_i), (3) combining facts and values into an overall value $v(a)$ (e.g. with an additive model¹), and (4) performing a

¹ As an example, the additive model for calculating the overall value of an alternative is presented:

$$v(a) = \sum_{i=1}^m w_i v_i(a_i)$$

where: $v(a)$ = overall value of alternative a (the value becomes a utility under risk)

a_i = attribute level of alternative a for attribute i (prediction)

$v_i(a_i)$ = value for attribute i of alternative a ; bounded to [0,1] (preference)

w_i = weight (or scaling constant) of attribute i , and sum of w_i equals 1 (preference)

Understanding this additive model is rather intuitive. However, to be used, three strong axioms should be verified (Eisenführ, F., Weber, M., Langer, T., 2010. Rational Decision Making. Springer Berlin Heidelberg, Germany, Berlin Heidelberg, Germany.):

- Simple preference independence (the preference of one objective does not depend on other objectives),

sensitivity analysis. Step (2) comprises weight w_i elicitation, among other preference parameters that are used for the decision model. Weight elicitation is a stepping-stone of individual preference construction, as individuals explore and formulate possibly new opinions about their own values, and clarify and/ or change their opinions (Belton and Elder, 1994) with regard to the relative importance of the objectives to be achieved. The decision analysts need to find a balance between (1) deepening the information provided to enhance understanding but avoiding biasing, (2) avoiding random errors but not introducing systematic errors, and (3) emphasizing meaningful aspects and lowering non-relevant aspects within the context (Fischhoff, 2006). Finding a balance, particularly for point (1), is critical when the participants are unfamiliar with at least some aspects of the complex decision problem, and need to learn.

1.2.3. Learning in adulthood

What is meant by learning? Learning is both a process and the by-product resulting from this process (Merriam et al., 2007). Learning as a process is designed to acquire information, assimilate it, and potentially change values. The learning process combines experiences and contextual, cognitive, and emotional influences in order to acquire, enhance, or potentially provoke these changes. Learning as a (by-)product constitutes a change in the individual's "knowledge, skills, values, and worldviews" (Merriam et al., 2007, p.277). Note that learning does not necessarily require being taught: learning can be incidental, i.e. an unintentional by-product of which individuals become aware only after a specific experience. In this paper, we focus on learning as an individual change in knowledge and values, because knowledge of complex problems is a precondition for a potential change of values.

Various theories of learning exist, some with a constructivist orientation, such as the Self Directed Learning theory (Garrison, 1997) and the Transformational Learning theory (Mezirow, 2000). In this constructive framework, learning occurs when people actively give meaning to their experience, individually or in social interaction. Learning is enhanced when the learner experiences cognitive dissonance (Adcock, 2012): a conflict created by the difference between what the learner already knows or believes and the new information. Cognitive dissonance induces the need for the learners to adapt their knowledge, mental model or values to the new experience. The Self Directed Learning theory emphasizes the

-
- Mutual preferential independence (generalization of the simple preference independence for any subset of the objective set),
 - Difference independence, or compensation (the additional value attached to an improvement of the objective does not depend on other objectives).

importance of motivation, as this triggers entering the learning task. After entering the learning task, self-monitoring (learner's responsibility) and self-management (learner's control) interact to finally direct learning (Garrison, 1997). The Transformational Learning theory highlights the role of critical reflection in learning: either by thinking about the experience (content reflection), by thinking about ways to deal with the experience (process reflection) or by re-examining social constructs established over the long-term (premise reflection) (Mezirow, 2000).

1.2.4. Learning to construct preferences

Given the assumption of preference construction and the theoretical background on adult learning, it appears natural to consider preference construction as learning. In fact, many decision analysts consider MCDA as a learning process (Gregory et al., 2012; Hämäläinen et al., 2001; Karjalainen et al., 2013; Marttunen and Hämäläinen, 2008). More generally, a research question has long asked whether Operational Research methods are effective learning tools (e.g. Belton and Elder, 1994; Lane, 1992; Monks et al., 2016; Thompson et al., 2016).

In the present paper, based on Transformational Learning theory (Mezirow, 2000), we propose to formalize preference construction as a learning process at individual level (Tab. 1). We suggest that Mezirow's content reflection, applied to the construction of reliable preferences, is twofold. First, one needs *factual learning*, i.e. acquiring data about the decision problem, and structuring it so that it becomes useful information for decision-making. Factual learning can reinforce a pre-existing mental model or induce a change in it; the mental model is defined here as the internal representation of the problem to be solved. Second, one needs *preference learning*, i.e. confirming or changing values relating to the various aspects of the problem, both objectives and alternatives in our case. Finally, Mezirow's process reflection, or *process learning*, could be a by-product of preference construction: the individual could reflect on the elicitation process used in MCDA and learn about a general way to solve a complex problem.

Table 1: Twofold adaptation of Mezirow's (2000) content reflection to factual and preference learning by individuals for application in MCDA.

2. Preference learning

How one values the various elements (objectives and alternatives) of the problem

Strengthen	Change
------------	--------

1. Factual learning How one acquires information and structures the problem: mental model (MM) (objectives, alternatives, predictions, relations between elements)	Same mental model	An individual acquires information that makes explicit his/her pre-existing mental model, and this reinforces his/her pre-existing preferences.	An individual acquires information that makes explicit and clarifies his/her pre-existing mental model; but in the light of this more conscious understanding of the situation, he/she values things differently.
	Different mental model	An individual acquires information that modifies his/her pre-existing model, or that generates a new, mental model; but this new understanding of the situation does not modify his/her pre-existing preferences (initially built on heuristics).	An individual acquires information that modifies his/her pre-existing model, or generates a new, mental model; and in the light of this new understanding of the situation, he/she values things differently.

In addition, learning theories have practical developments which could benefit decision analysts. In particular, decision analysts could enhance learning, i.e. preference construction, by offering individuals an experience that triggers motivation, provides knowledge to understand the decision problem, and enables reflective thinking, e.g. by creating situations of cognitive dissonance (Plass et al., 2015).

1.2.5. Gamification enhances learning to construct preferences

Attempts have been made to increase public participation in MCDA using web-supported technologies, but these have been criticized (as shown in Section 1.1). The lack of interaction between analyst and respondent may increase the risk of eliciting unreliable preferences. We propose to compensate the missing guidance of the experienced decision analyst by “gamifying” online preference elicitation.

Gamification means “using elements of game design in non-game contexts, products and services to motivate desired behaviors” (Deterding, 2012). Gamification and game-based learning are designed to trigger psychological processes which in turn have behavioral consequences, such as learning (Lane, 1995; Plass et al., 2015), or increased engagement (Hamari et al., 2014; Seaborn and Fels, 2015). A recent paper reviewing 14 publications on gamified surveys (from survey and marketing research) reports that gamification has positive effects, e.g. on peoples’ participation, the survey time, or their willingness to answer demanding questions (Keusch and Zhang, 2015). However, these effects are not systematic,

and the experimental designs make comparison and generalization difficult. There are no ready-to-use solutions for gamification (Morschheuser et al., 2017): any gamification design is tailored to suit the specificities of the project. The gamification of our survey is described in Section 2. Most of the recent literature on gamification calls for proper evaluation of the gamification benefits (e.g. Hamari et al., 2014; Seaborn and Fels, 2015).

Before proposing a suitable concept, we reviewed the use of gamification (and serious games) for environmental issues (Aubert et al., 2018). No existing gamification (or serious games) allowed for MCDA-compatible preference elicitation. Nevertheless, we concluded that gamification would be a promising approach to provide information for better understanding the decision problem, to create situations of cognitive dissonance, and to trigger engagement, thus potentially allowing reliable online preference elicitation by many individuals. To this end, we conceptualized and developed a gamified online preference elicitation survey designed to promote (1) factual learning, (2) preference learning, and to provide (3) a positive, engaging experience. In particular, we tested whether the prototype of the gamified survey verified the following hypotheses:

(H1) on factual learning:

H1a: the proposed learning loops enhance learning about the decision topic (wastewater management),

H1b: participants are aware of this factual learning,

H1c: factual learning is greater for novices and those who spend more time on the informative parts (descriptions of objectives and alternatives);

(H2) on preference learning:

H2a: the proposed learning loops enhance participants' learning about how they value objectives,

H2b: the proposed learning loops enhance participants' learning about how they value alternatives,

H2c: the proposed learning loops do not introduce undesired biases in preferences (note: this hypothesis is needed to check whether respondents showed systematic behavior to trick the system in order to reach consistency);

(H3) on positive experience:

H3a: the newly developed preference elicitation survey provides a positive experience,

H3b: this experience is more positive for participants answering the gamified survey compared to those answering the same survey without game elements (control).

2. The concept of the online preference elicitation survey

Our suggested gamification would be classified as “soft gamification” of a survey (Bailey et al., 2015), as it uses well-established MCDA methods and questionnaires for preference elicitation and incorporates game elements into them. The game elements are (1) learning loops, (2) a narrative with non-player characters, and (3) a friendly interface design.

The learning loops are based on cognitive dissonance: they could be seen as win-fail situations. If the respondent is facing a dissonance, he/she needs to evaluate his/her actions, and adapt his/her assumptions to solve the dissonance. Cognitive dissonance provokes a potential change of the mental model, i.e. learning. Cognitive dissonance is found in many learning theories, e.g. the Kolb learning cycle, and is sometimes referred to as feedback learning. It is also one of the key elements of game-based learning (Plass et al., 2015). In the survey, we developed learning loops at two levels (Fig. 1): a learning loop on objectives and another on alternatives. We needed to avoid judging the values of the participant. Therefore, he/she sets his/her own references using various elicitation methods. The preferences obtained with these methods are compared. If they are consistent, this shows that the respondent knows about his/her preferences and how to express them. If the preferences are inconsistent, this shows either that the respondents’ preferences have not yet been constructed or that he/she does not know how to express them with our methods (see next paragraph regarding consistency of methods). This feedback element has so far generally been missing in online MCDA surveys. We feel that this solution is appropriate for MCDA gamification as it can mimic the role of the decision analyst by giving feedback on consistency in face-to-face interviews. Also, it complies with recommendations to move towards a procedure that reconciles different methodological approaches (Anderson and Clemen, 2013; Hobbs and Horn, 1997).

Figure 1: The online MCDA preference elicitation survey flow. There are two levels of learning loops: the learning loop on objectives (Step 2 and sub-steps, light grey boxes) is embedded in the learning loop on alternatives (black boxes).

More specifically, the learning loop on objectives is based on two widely used methods of weight elicitation: trade-off and swing (e.g. Eisenführ et al., 2010). Here, in Step 2c (Fig. 1, grey boxes), the participant is confronted with the results of the two weight elicitation procedures he/she has just performed in Steps 2a (swing) and 2b (trade-off). We acknowledge that in practice variance between weight elicitation methods has been reported (Fischer, 1995; Lienert et al., 2016; Pöyhönen and Hämäläinen, 2001; Weber and Borcherting, 1993). In addition, the ranking can be different while the weights can be of the same order of magnitude. Thus, we aim at soft consistency. If the ranking of the objectives is the same and the differences in the weights assigned to the objectives are below 10% (i.e. <0.1 if sum of weights equals 1), we consider the results to be consistent; we provide positive feedback and the survey can continue. If the ranking of the objectives is the same and the differences in the weights assigned to them are above 10%, positive feedback is produced, together with a feedback warning that the weights assigned with the two methods differ greatly for the given objectives. In this case, the participant is given the chance to re-consider

the weights, i.e. his/her preferences as regards the importance of the objectives. He/she can decide which weight elicitation method provided the numerical values closest to his/her preferences or can choose to repeat either of the methods. If the ranking of the objectives is inconsistent, the participant failed and may repeat either the weights preference loop or the method that provided the ranking which does not fit his/her preference. Note that after three preference elicitation trials (initial swing, initial trade-off and repetition of one of the methods), the participant has the option to move on in the survey to avoid tiring and drop-out. In this case, the player needs to specify which weights should be used for the next steps. The swing method is composed of two steps: first, the participant ranks hypothetical alternatives, and second, he/she assigns a score from 100 (most preferred) to 0 (least preferred) (see Appendix 1a). The trade-off method comprises two steps repeated $n-1$ times (with n the total number of objectives): first, the participant makes a preference judgment between two alternatives of two objectives, one being at the worst level and the other at the best level, and second, he/she adjusts the fulfillment of the least preferred alternative to reach indifference (see Appendix 1b).

The learning loop on alternatives is a consistency check based on the ranking of alternatives: holistic ranking vs. MCDA ranking. In Step 1, the participant is asked to perform an initial holistic ranking of all alternatives. Holistic ranking is not generally recommended in the MCDA literature (Keeney and Raiffa, 1976), but is widely used in practice. It involves the decision alternatives being directly presented to the participants, who then rank them from best to worst. This approach has been criticized because, for complex alternatives that involve different objectives, the ranking may be based on decision heuristics rather than on a thorough consideration of the participant's real values and his/her preferences concerning the trade-offs that have to be made between different objectives (i.e. value-focused thinking, Keeney, 1996). Then, after the weights preference elicitation loop (learning loop on objectives), an MCDA ranking of the alternatives is obtained based on the overall performance of each alternative: it is calculated using the weighted sum aggregation model (see Footnote 1). In Step 4, the participant is confronted with the results of the two rankings, the holistic ranking and the MCDA ranking based on weight elicitation. If the participant is consistent (Step 4A), which could happen in particular if he/she is knowledgeable about the decision issue, we ask whether he/she is confident that the result reflects what he/she thinks about the decision, or whether he/she wants to repeat any step. If the participant is inconsistent (Step 4B), we provide a graphical representation of the prediction matrix (i.e. the attribute levels of each alternative, a_j). We subsequently ask for an informed holistic ranking of alternatives based on better information about the performance of the alternatives

(Step 5). If this informed holistic ranking is consistent with the MCDA ranking, we ask the participant whether he/she is confident that the result reflects what he/she thinks about the decision, or whether he/she wants to repeat any step. If the ranking is inconsistent, the participant can decide between repeating the whole weight elicitation loop, following an individual method, or again repeating the holistic ranking (Fig. 1). If the participant is in this situation of inconsistency for a third time, a fourth option appears: he/she can stop the survey.

The game elements offering motivational affordances constitute a narrative to give an additional mission to accomplish, and thus to increase motivation, and non-player characters to give feedback and encouragement, replacing the absent decision analyst. Finally, the design of the interface is attractive and intuitive.

The prototype was developed using data on sustainable wastewater infrastructure planning in a Swiss case from an earlier SWIP project: the hierarchy of objectives, the set of alternatives, and the prediction matrix are publicly available (Lienert et al., 2016, and Supplementary Information). We invite the reader interested in the wastewater decision to consult the original SWIP publications. For the prototype testing aimed at a proof of concept, the decision problem was simplified by reducing it to four objectives and four alternatives. The selected objectives were: highly flexible system adaptation, high recovery of phosphate, low time demand for end users, and low annual cost. The selected wastewater management alternatives were: (1) status quo – a centralized and relatively good wastewater treatment system (A8a), (2) centralized high tech – a greatly improved centralized system (A2), (3) decentralized low tech – the community provides a minimal service and households are responsible for a cheap decentralized wastewater system (A5), and (4) decentralized with nutrient recovery – cooperative and private responsibility of the decentralized wastewater system allowing nutrient recovery for agriculture (A7). The descriptions used for this experiment are presented in Appendix 2.

The development of the survey was outsourced after a competitive selection. The selected company, Takomat GmbH², developed a suitable app. The prototype used for the experimental proof of concept was developed solely for iOS devices.³ It requires a Wi-Fi connection for data exchange and storage on the server hosted by the company. The authors received the raw database on demand.

² Köln, Germany, www.takomat-games.com, retrieved on 22.05.2018

³ The prototype is accessible on the iTunes Store: look for the Eawag survey App. It is publicly available, but tokens are required to log into the gamified and control survey. These tokens are available from the main author upon request.

3. Data & Methods

3.1. A mixed experimental design

To test the prototype experimentally, a more conventional survey was developed as a control. The logic of the control survey was identical to the gamified survey but without the game elements included for motivational affordances (no narrative, no non-player character). The control survey also had a more sober design. See Fig. 2A as an example. The development of the control survey was outsourced to the same company as the gamified survey. The data were stored in the same database.

Figure 2: Screenshots from the same step in the gamified survey (left picture, framed in purple) and control survey (right picture, framed with a dashed green line).

The experiment consists of comparing a gamified and a control treatment. All answers given by the participants during the respective treatments were recorded: e.g. number of learning loops, chosen elicitation method (if repetition), time spent, and the results of each elicitation.

Both treatments were preceded and followed by ante- and post-questionnaires respectively. The ante-questionnaire included socio-demographic questions and an initial knowledge test on wastewater management. The post-questionnaire included a repetition of the knowledge test and feedback questions about the experience (see Appendix 3).

In addition, qualitative data were collected. First, all participants were invited to provide comments in an open box at the end of the survey, or email back written comments. Second, we organized a workshop entitled “testing a gamified interview tool for MCDA” targeting game design experts, i.e. game design students from Zurich University of the Arts (<http://gamedesign.zhdk.ch/>, retrieved on 21.03.2018). The workshop was facilitated by the first author, assisted by a scientific collaborator. To begin with, participants tested either the gamified or the control survey. Directly following the test session, they were asked to individually fill in a paper questionnaire with three tasks: (1) to write down the first two adjectives that came to their mind to describe their experience of the prototype; (2) to rate six aspects of the experience (e.g. difficulty, motivation, fun, etc.) on 5-Likert scales; and (3) from a list of 11 game elements (adapted from Hamari et al., 2014) to tick those they had recognized, to comment its effect in each case and to rate whether the use of the element was successful (material in Appendix 5a) on a 5-Likert scale. After a 15-minute break, tasks (1) and (3) were collectively discussed in an audio-recorded session, confronting the comments of the participants who tested the control with those who tested the gamified survey.

3.2. Measures used to test each hypothesis

3.2.1. Measures to test factual learning hypotheses (H1)

H1a (the proposed learning loops enhance learning about the decision topic) could be tested thanks to the knowledge test in the ante- and post- questionnaires. This test comprised four questions, one for each objective (see Appendix 3 for the questions). One point is given for each correct answer. The initial and final knowledge scores vary from 0 (no correct answer) to 4 (all answers correct). Factual learning occurred if the score of the final knowledge test was higher than that of the initial knowledge test, i.e. if the score difference was positive. H1b (participants are aware of this factual learning) and H1c (factual learning is higher for novices and those who spend more time on the informative parts) are presented in Appendix 6.

3.2.2. Measures to test the preference learning hypotheses (H2)

H2a (the proposed learning loops enhance participants' learning about how they value objectives) and H2b (the proposed learning loops enhance participants' learning about how they value alternatives) were tested on the basis of the following assumption: when the participants have learnt about their preferences, they are able to state them consistently

using different methods. Thus, we used consistency between the two methods composing the respective learning loops (defined with the rules mentioned in Section 2) as a proxy for preference learning. We calculated the proportions of respondents who reached consistency in the objective and alternative learning loops, for H2a and H2b respectively, and compared the two treatments.

In addition, the literature suggests that learning may be defined as a potential change in values (Merriam et al., 2007). Thus, we tested if the ranking of each alternative changed between the initial elicitation (done before participants saw the objectives) and the final one: we calculated a rank difference defined as the initial rank minus the final rank of that alternative ($-3 < \text{rank difference} < 3$). We particularly looked at the status quo alternative: we assumed that preference learning made participants move away from the status quo bias.

To test H2c (the proposed learning loops do not introduce undesired biases in preferences), we identified the obvious ways to trick the system in order to reach consistency (necessarily used by the author during the prototype development). Occurrences of these specific patterns would validate this hypothesis. For the objective learning loop, the patterns are the following: giving equal weights to all objectives, or giving all the weight to a single objective (and being indifferent about the three others). For the alternative learning loop, the pattern is the following: choosing to repeat the holistic ranking and feeding into it the ranking obtained with the MCDA valuation.

3.2.3. Measures to test the positive experience hypotheses (H3)

Hypothesis (H3) on the positive experience was investigated using different measures. First, we assumed that if the motivational affordances of the gamified treatment were effective, the participants would be more willing to reach consistency, and would be willing to undergo more learning loops, than those answering the control treatment. We thus counted the number of learning loops.

Second, participants assessed five statements about their experience on 5-Likert scales from “I strongly disagree” to “I strongly agree” (e.g. “The questionnaire kept me motivated while filling it in”, see Appendix 3). A specific question assessed the perceived difficulty, which was further investigated by asking participants to tick the reasons for the difficulties they met from a list of items (Section 4.3, Tab. A7, Appendix).

3.3. Data collection

We wanted a reasonably homogeneous sample, as the literature on gaming reports age and gender effects (e.g. Koivisto and Hamari, 2014). We targeted students (from BSc/Ba to PhD) and young academics (postdocs). Participants were recruited in our research

institute (Eawag: www.eawag.ch, retrieved on 21.03.2018), the ETH Zurich (www.ethz.ch, retrieved on 21.03.2018), and the University of Zurich (www.uzh.ch, retrieved on 21.03.2018) through emailing or public advertisement on the student platform (www.marktplatz.uzh.ch, retrieved on 21.03.2018). Participants were given an incentive, following the standards of the University of Zurich. Every second male and female participant was assigned to the gamified version to ensure the same gender balance in the control and gamified groups.

In total, 107 participants answered the survey from start to end, of which 53 answered the control and 54 the gamified survey: 57 were male and 50 female, equally distributed among the survey type (female: 24 control, 26 gamified; male: 29 control, 28 gamified).

3.4. Data analysis

The statistical analyses were performed using R project for statistical computing (R Development Core Team, 2017). The normality of the data was checked using the Shapiro-Wilk normality test. Accordingly, and depending on the nature of the variables, we evaluated the differences between groups using student t-tests, Wilcoxon tests, and binomial tests. Further details are given in Section 4.

The qualitative data complemented the quantitative results and helped to provide explanations. We looked for two themes in the qualitative data: (1) learning, with the classification proposed in the introduction: factual, preference, and process learning; and (2) the experience provided by the survey. To address the latter, we categorized the qualitative comments into the three aspects of the Self Determination theory which has already been applied to games and media (Rigby, 2014; Ryan, 2017; Ryan et al., 2006; Tamborini et al., 2010). According to this theory, a game or media is successful when it motivates the participants, and this motivation arises when three fundamental psychological needs are fulfilled: competence (or mastery), autonomy, and relatedness.

4. Results

4.1. Factual learning (H1)

H1a (the proposed learning loops enhance learning about the decision topic) was verified. Overall, participants had a statistically significantly higher knowledge score in the post-questionnaire compared to that reported in the ante-questionnaire (Tab. 2), independently of the survey type. Some qualitative comments also supported the occurrence of factual learning (Tab. 4, Row 1).

Table 2: Results of factual learning. Initial and final knowledge scores: number of correctly answered knowledge questions in the ante- and post-questionnaires respectively; values can range from 0 to 4. Score difference = final – initial, values can range from -4 to 4. Perceived factual learning, values can range from 1 (not at all better understanding) to 5 (much better understanding). Med=Median, n.s.= non-significant.

		Initial knowledge score [0, 4]			Final knowledge score [0, 4]			Score difference [-4, 4]			Perceived factual learning [1, 5]		
		Mean	SD	Med	Mean	SD	Med	Mean	SD	Med	Mean	SD	Med
Overall	N=107	1.54	0.95	2.00	2.95	0.97	3.00	1.41	1.08	1.00	3.43	0.89	4.00
Gamified	N=54	1.80	1.09	2.00	3.02	0.96	3.00	1.22	1.13	1.00	3.63	0.76	4.00
Control	N=53	1.28	0.72	1.00	2.89	0.97	3.00	1.60	1.01	2.00	3.23	0.97	3.00

Results of the Wilcoxon tests:

Between initial and final overall knowledge scores: W=1800, p-value=2.2e-16
 Between score difference from gamified and control surveys: W=1718.5, p-value=6.4e-2 (n.s.)
 Between initial knowledge score from gamified and control surveys: W=1056, p-value=1.4e-2
 Between perceived learning from gamified and control surveys: W=1112, p-value=3.0e-2

4.2. Preference learning (H2)

H2a (the proposed learning loops enhance participants' learning about how they value objectives) was verified. A binomial test (probability of success 61%, p-value=3.0e-2) showed that most people (65 persons out of 107) ended the survey statistically significantly with weights that were consistent between the two elicitation methods, swing and trade-off. The test indicated that this difference was not random.

H2b (the proposed learning loops enhance participants' learning about how they value alternatives) was partially verified. Overall, 58 participants finished the survey with holistic ranking of alternatives which differed from the ranking obtained from the MCDA valuation (i.e. they had an inconsistent ranking of alternatives), and 49 finished with a consistent ranking. A binomial test (probability of success 46%, p-value=4.4e-1, n.s.) indicated that the consistent ranking of alternatives could be a random phenomenon. However, a Wilcoxon test between the initial and the final rankings of alternatives resulted in statistically highly significant differences (W=2623, p-value=3.9e-09): participants changed their views about the alternatives. Half of the participants moved away from the status quo alternative (rank difference of the status quo alternative: mean=-0.6, SD=0.9): the binomial test implied that moving away from the status quo could be a random phenomenon (probability of success 50%, p-value=1, n.s.).

H2c (the proposed learning loops do not introduce undesired biases in preferences) was verified. None of the typical patterns used to reach weight consistency was observed. No alternative consistency was significantly noted (see H2b). These results validated H2c.

Some qualitative feedback (Tab. 4, Row 2) supported the occurrence of preference learning.

4.3. Positive, motivating experience (H3)

H3a (the newly developed preference elicitation survey provides a positive experience), and H3b (this experience is more positive for participants answering the gamified survey compared to those answering the same survey without game elements (control)) were partially verified.

There was no statistically significant difference in the number of learning loops between the control and gamified surveys (Wilcoxon test, $W=1495.5$, $p\text{-value}=6.1e-1$, n.s.): on average, the participants repeated two learning loops. Overall, the perception of the survey was neutral to positive, both for the gamified survey and the control (Tab. 3). According to the self-reported rating on the perception of the surveys as a “motivating tool”, “fun experience”, “worth recommending to others”, “worthwhile use of time”, and “usefulness of answers” (Tab. 3), there were no statistically significant differences between the control and gamified surveys. However, the medians indicated that the gamified survey tended to be perceived better on all of these questions than the control.

510 Table 3: Results of the post-questionnaire on 5-Likert scales. SD=standard deviation; Med=median; n.s.= non-significant

Survey perceived as...		Motivating tool			Fun experience			Worth recommending to others			Worthwhile use of time			Usefulness of answers			Ease/ difficulty		
		Mean	SD	Med	Mean	SD	Med	Mean	SD	Med	Mean	SD	Med	Mean	SD	Med	Mean	SD	Med
Overall	N=107	3.34	0.92	4.00	3.15	1.01	3.00	3.46	0.90	4.00	3.38	0.92	3.00	3.44	0.81	4.00	3.48	0.89	3.00
Gamified	N=54	3.42	0.89	4.00	3.24	1.01	4.00	3.62	0.90	4.00	3.45	0.80	3.00	3.54	0.87	4.00	3.56	0.82	3.00
Control	N=53	3.26	0.96	3.00	3.06	1.01	3.00	3.30	0.87	3.00	3.30	1.03	3.00	3.43	0.75	4.00	3.40	0.97	4.00
Results of Wilcoxon tests between gamified and control survey:																			
	W	1285.5			1293			1139			1338.5			1407			1357.5		
	p-value	4.2e-1 (n.s.)			3.7e-1 (n.s.)			7.4e-2 (n.s.)			6.6e-1 (n.s.)			9.9e-1 (n.s.)			6.3e-1 (n.s.)		

There was no statistically significant difference regarding the perceived difficulty of the survey (Tab. 3, last column) but the medians indicate that the control would be perceived as difficult and the gamified survey as neutral.

According to the feedback questions on the survey difficulty (Tab. A7, Appendix), 21.5% of participants reported a lack of sufficient information (on the topic), independently of survey type. In contrast, 16% of the participants reported that there was too much information, and statistically significantly more participants who answered the control survey reported that they had received too much information (23%) compared to those who answered the gamified survey (7%). The second part of the trade-off questions (when adjusting the score until they were of equal value) was often reported as difficult in both surveys. Equally, the scoring part of the swing method was reported as difficult.

The qualitative results supported the quantitative ones. Some respondents reported that there was too much textual information, while others reported that they had received insufficient information to make a decision (about the consequences of the values assigned to the attributes in the weighting tasks, and the description of alternatives). The trade-off method was reported as being difficult twice in the voluntary written comments, while the swing method was not explicitly mentioned as being difficult in this open feedback. The difficulty in the scoring part of the swing method according to the ticked feedback answers (Tab. A7, Appendix) may be due to a usability issue, as many respondents reported technical issues in answering this particular part of the survey. In addition, overall, the qualitative feedback on the “learning loop”, where the aim was to achieve consistency between different methods, was negative (Tab. 4). However, a positive general perception of the surveys was also reported (Tab. 4), in particular as regards the game elements included to provide motivational affordances (narrative and non-player characters in the gamified survey). Students from the game design section of Zurich University of the Arts were much more critical towards the proposed gamification. In the first workshop task, when they wrote down the first two adjectives that came to mind to describe their experience, about three quarters of the adjectives were negative and only one quarter were somewhat positive (Appendix 5b). Some of their comments are reported in the discussion as inputs for improvements.

Table 4: Report of some qualitative comments supporting the various hypotheses

Hypotheses	Qualitative data
H1a factual learning occurrence and H1b awareness	e.g. “I learned quite something!”, “We know nothing about [wastewater management] and we start to learn about this”, “You learn what a centralized or a decentralized wastewater system is at least”, “I heard some stuff that I

	didn't know before"
H2 preference learning occurrence	e.g. "Because somehow I learned something in between, I changed my opinion", "In the first phase where you learn things, you accept that more and have a different mind-set on that"
H3 perception of the learning loop	e.g. "confused [by the inconsistency]", "close to abort", "gave up [after] not know[ing] what to shift anymore in which direction"
H3 perception of the control survey	e.g. "was a fun experience"
H3 perception of the gamified survey	e.g. "nice survey!", "it encourages careful reading and answering", "nice job!", "best survey design so far", "this is the best survey I have ever completed"
H3 perception of the game elements included to provide motivational affordances	e.g. "there is a story and it makes it easier to answer", "I thought it looked really nice and it is more motivating to fill in such a survey when it is presented like an interactive game", "the two characters: LOVE THEM!, [...] it is not just more entertaining but it also feels "easier" to answer the questions"

5. Discussion

5.1. The experiment provides a positive proof of concept

Our experiment clearly confirmed that factual learning on wastewater management and preference learning about the objectives did in fact occur (H1a and H2a). The proposed learning loops not only increased the participants' knowledge about the facts (H1a) but also helped them to learn about their own preferences concerning their valuation of the importance of the objectives. After several preference elicitation loops, statistically significantly more participants achieved consistent weights between the two weight elicitation methods, namely swing and trade-off, than at the beginning of the survey (H2a; Step 2c in Fig. 1). Thus the proposed concept based on learning loops using well-established methods of the MCDA field proved to be a powerful tool for online preference elicitation. It enables weight preferences to be collected from a large number of informed laypeople without a facilitator. Our results contribute to the debate on online preference elicitation. They support the following position: with carefully designed surveys, aimed to enhance reflective thinking and learning, online preference elicitation from many citizens is possible. Others had previously claimed this by using "pathway surveys" (Gregory et al., 1997; Gregory et al., 2016). These surveys combine tutorials, value positions and trade-offs in an iterative process. They are meant to represent the range of deliberative rationales encountered in the population. Our concept differs from pathway surveys mostly because it does not require us to assume the "pathways" or lines of thought found among the general public, and it relies on

the well-established MCDA elicitation methods. Other hypotheses on factual and preference learning are discussed below.

We are pleased that the clear rejection of hypothesis H2c indicates that the proposed concept does not introduce undesired biases during preference construction. None of the participants played with the game design or created strategies by interpreting the rules, which is actually expected in the game design literature (e.g. Sicart, 2013). For our application this means that none of the participants wanted to be consistent for the mere sake of being consistent, independently of their preferences. This indicates that the survey was taken seriously, despite the gamification, and that participants answered according to their own preferences. This may have been due to the “soft gamification” concept: the survey remained “serious”, asking participants to answer according to their own preferences, and included only a few game elements (Bailey et al., 2015). These introduced elements obviously did not jeopardize the validity of the collected preference data. This further supports our conclusion that the proposed concept would enable reliable weight preferences for MCDA to be collected from informed citizens.

5.2. Suggestions to improve future work

5.2.1. On gamification

As mentioned in Section 4, the game design students were critical about the gamification. We discuss their comments below, organized into three sub-themes: competence (or mastery), autonomy, and relatedness (Self-Determination theory, e.g. Ryan et al., 2006). We suggest means to improve the gamification. Implementing the suggestions should increase the significance of our results in terms of a positive experience (hypothesis H3).

For some participants, the proposed gamification caused a feeling of lack of competence, which was mirrored in the adjectives the game design students used after the survey (e.g. “unqualified”, “confused”, “frustrating”, “too complex”), and their feedback in the discussion (e.g. “At first, I was quite open to fill out the survey, quite happy, and then the survey told me: You didn’t do this right, you’re an idiot! And then it started to feel like work.”). Some participants obviously felt unsuccessful or ineffective. This could relate to the learning loops, which seem to have been too difficult to handle: participants had difficulties in understanding how to reach consistency, particularly for the learning loop on alternatives (between the “informed holistic ranking of alternatives” in Step 5 (Fig. 1) and the outcome of the MCDA (ranking of alternatives based on their weight preferences)). This interpretation is

consolidated by the quantitative results from all respondents, where only about half of the participants achieved this consistency between the holistic ranking and the MCDA ranking in Step 6 (hypothesis H2b was only partially verified).

Additionally, the quantity of new information to remember and mobilize may also have undermined the feeling of competence. To increase the feeling of competence (mastery) in future developments, we offer three proposals. They are based on the idea that to keep engaged – or in the “flow” – one should be challenged, but the challenging task should not be overwhelmingly difficult; it should seem reachable (Csikszentmihalyi, 2000). First, the rules defining consistency in the learning loops could be adapted in many possible ways. Relaxing the consistency rules would acknowledge that most human beings are unable (or unwilling) to behave completely rationally. Although the MCDA process is meant to facilitate rational thinking, it might be too challenging for untrained laypeople. For instance, for the learning loop on preferences regarding the importance of objectives (weight elicitation; Step 2c, Fig. 1), the 10% buffer threshold for the weight consistency check between two different weight elicitation methods could be adjusted. The adjustment could be based on a sensitivity analysis of the weights if this was available for the decision context. Additionally, for the learning loop on preferences regarding alternatives (Steps 4 and 6, Fig. 1), the consistency check could rely less strongly on the rankings of alternatives. Instead, the ranking of alternatives from MCDA weight elicitation would be equal to the holistic ranking (Steps 1 and 5) if the overall MCDA values of the alternatives (Step 3) are within a pre-defined buffer range. Such changes would make the consistency rules more flexible so that consistency would be easier to reach, which would increase the feeling of competency.

Second, better adapted feedback on the “how to” aspect could be provided, even though some studies report that feedback can be perceived as a burden, contrary to its definition (Ryan, 2017 and references therein). Useful feedback should be developed in such a way that the cognitive load of the survey is reduced, e.g. by clarifying how to elicit one’s own preferences and by giving feedback on the meaning of the elicited preferences. As one example, feedback on the rating of objectives using swing weight elicitation might be: “The number you just gave means that improving objective X from its worst to its best state is only one third as important to you than improving objective Y from its worst to its best state.” However, such personalized feedback requires relatively complex interactive programming.

Third, the software could enhance multimodal learning using interactive multimedia. It could include various information representations (e.g. text and images), be more interactive and personalized (see the example above), and involve different senses (e.g. eyes and ears)

(Massaro, 2012; Niegemann and Heidig, 2012). This could facilitate new information processing, in turn increasing the feeling of competence/mastery.

For other participants, the proposed gamification caused a feeling of lack of autonomy, which was mirrored in the choice of adjectives of the game design students (e.g. “forced”, “prejudiced (advertising something)”, “click-marathon”). Participants had little choice in deciding on the survey path. This could relate to the excessively strict definitions of consistency discussed above (unsuccessful participants felt controlled), some usability issues (e.g. the sliders could be improved from sliding the cursor on the scale to touching the scale), and the fixed structure of the survey. Our first proposed survey concept did not create enough opportunities from which the participant could choose freely. Future developments should offer: (1) more options, e.g. after an initial learning loop (a forced cognitive dissonance) there could be an option to (1a) manually adjust the elicited preferences, or (1b) confirm that although consistency was not reached, one of the elicited preferences reflected the users’ preferences. Further developments should also allow (2) the consistency rules (as described in the above paragraph on competence) to be reviewed, and (3) the usability to be improved (e.g. go back button, better sliders). Some participants suggested that they should have had the opportunity to add some objectives that are important to them, which would be a further argument for introducing a stepwise process (as discussed in Section 5.2.2 below).

Other participants felt a lack of relatedness which was reflected in their comments (e.g. “the story does not engage [the participants]”, “you totally lose the connection”, “[the non-player characters] felt like placeholders”). These participants did not feel supported by or connected to others, nor did the non-player characters increase the feeling of connectedness. A review of gamified surveys in marketing research also reported that including a narrative was harmful because it disconnected respondents from the survey topic and increased drop-out (Keusch and Zhang, 2015). To address this frustration, future developments could offer: (1) a more integrated story with a narrative more connected to the decision topic, and (2) room for exchanges with people having different opinions (either other participants or the non-player characters).

Finally, some of the game design students actually questioned the decision to gamify the survey, as Keusch and Zhang (2015) did at the end of their review: “[are] surveys even supposed to be fun”? In our opinion, this conclusion is based on wrong assumptions: “fun” should not be the aim of gamification (Rigby, 2014), but gamification should be seen as a means to engage people thanks to the motivational affordances, i.e. the game elements (Hamari et al., 2014). In our case, given the validation of our hypotheses on factual and

preference learning, as well as the overall rather positive experience despite the discussed criticism, we were able to confirm that gamification was useful. Even though gamification might not be fun in all cases, it would be worth implementing the suggested improvements to increase the positive perception of a preference elicitation survey.

5.2.2. On the prototype survey concept

The concept of this prototype enabled factual and preference learning. However, a participant of the workshop raised the following issue: “Normally, when you try to form an opinion on something, you try to think about it, sleep on it, or really look at different sources of information. In this [survey] case, you only have this information, someone who forces you to form an opinion about it immediately.” This participant continued with: “for me it’s just too quick, [...] I can’t verify it because I am in this app and I have to decide now”. We think that this game design student has raised a major issue concerning preference elicitation for complex decisions. She is fully in line with Gregory et al. (2012, p. 26) “People need time, to think and to feel, in order to come up with a relatively stable understanding of their own preferences. Participants in a decision process should be allowed to reflect on choices outside of the meeting room environment”. For this reason, it might be worthwhile to consider a several stage preference elicitation process. For instance, a first step could be dedicated to factual context-specific learning, followed some days later by a second step dedicated to preference learning.

Other workshop participants felt that the survey was “prejudiced (advertising something)”, further explaining that the choice of the objectives was not neutral and that this influenced them too strongly. This might be due to the selection of only four objectives for the prototype for efficiency reasons: not all of the most important trade-offs between objectives were presented here. This conclusion is backed up by critical qualitative feedback from participants from Eawag experienced in wastewater management.

Alternatively, this criticism may raise another point: broad citizen participation in the preference elicitation phase could be insufficient. Citizens could claim the right to be involved earlier in the MCDA process, e.g. to co-construct the decision frame and objectives hierarchy. This would give them (1) more control over the decision-making process, and (2) more in-depth and a longer association with/relatedness to the decision-making process and topic. Both aspects are components of psychological theories: for instance, the Psychological Ownership theory (Pierce et al., 2001), which in turn impacts commitment; or the Self-Determination theory (Rigby, 2014), which in turn impacts the internalization of extrinsic motivation. Thus as one example of co-construction, developing an online tool for

building an objectives hierarchy based on citizen feedback would contribute to enhancing citizens' engagement in decision-making. Such a tool is currently being developed for wastewater management (Haag and Lienert, in preparation); it could be combined with the preference elicitation survey proposed here.

In addition, this first attempt to gamify preference elicitation focused exclusively on weight elicitation. Thus it assumed linear marginal value functions and the additive (linear) aggregation model under certainty (as presented in Footnote 1). However, these common simplifications in MAVT applications are not necessarily appropriate (Haag et al., 2018; Langhans et al., 2014; Reichert et al., 2015). Clearly, successive work could address the elicitation of other preference parameters for MCDA models, including the shape of single-attribute value functions, parameters of a non-additive aggregation model, and the decision-makers' attitude to risks.

Note that the survey, like any tool, may suffer from the media. The internet is not as democratic, accessible and flexible as many claim (Dillman et al., 2009; Merriam et al., 2007), although it can be seen to be developing quickly. Thus, online surveys may suffer from sampling errors which may be increased by further restrictions, such as the use of a specific operating system. It has also been well established in survey research that any additional technical issues (e.g. finding the right device, or downloading an app, or instable internet connections) lead to more drop-outs (Keusch and Zhang, 2015). These points are not relevant for the proof of a concept as in our case, but they need to be considered in any further development of such surveys, to ensure that they are as accessible as possible to a broad range of citizens.

5.2.3. On the experimental design

Our concept of enhanced learning and the experimental design comparing the initial and final states allowed us to verify this hypothesis with statistical significance. It is reassuring to know that the learning effect was measured in both surveys, as they included the learning loops which create cognitive dissonances. However, we did not observe any statistically significant differences between our control and the gamified survey regarding a more positive, motivating experience due to the introduction of game elements (hypotheses H3 were not confirmed). The effects of the motivational affordances (narrative and non-player characters) on the experience (H3) may have been too weak to overcome the limitations in control and autonomy emerging from the excessively strict learning loops (see discussion in Section 5.2.1.).

Although process learning was not investigated in the present study, it was brought up in some of the qualitative feedback which showed indications of a cognitive dissonance regarding the process: e.g. “I was rather fascinated by how my preferences/choices were interestingly inconsistent”, “What am I doing here?”. Other comments indicated reflective thinking about the decision-making process, e.g. highlighting the complexity of the trade-offs handled with the MCDA: “If I had a very good job, I would rather pay more [for wastewater management] and work less, but if I had time...”, or reporting a partial understanding of the survey process: “I did not really understand how the goals were ranked in the second option exercise”. Although our experimental design did not aim at testing process learning, we would like to highlight its potential for future work.

Furthermore, we would like to mention that despite the random distribution of the participants between the control and gamified surveys, we observed a bias: there was a significant difference between survey types in the initial knowledge scores. Without this being intended, the participants who received the gamified survey initially knew more about wastewater management than those who were given the control survey. This may contribute to explaining the non-significant results of the participants’ own awareness that factual learning had occurred (hypothesis H1b, Appendix 6) and offer possible explanatory factors for factual learning (which we had hypothesized would be higher for novices and those spending more time on the informative parts; hypothesis H1c). Also, the skewed selection procedure may have impacted the perception of the experience, as experts and novices were not challenged by the same tasks. In the educational sciences, this effect is named expertise reversal: instructional procedures facilitating learning for novices lose effectiveness when the level of expertise increases (Sweller, 2012). Better control of the participants recruiting and assignment to a treatment would benefit future work.

Finally, further work could focus on the effect of gamification on the participant’s engagement (as described for instance in Ryan, 2017; Ryan et al., 2006; Tamborini et al., 2010). The independent variables could be the type of survey (control or gamified) as well as individual characteristics. The dependent variables would need to express the expected psychological outcomes (engagement, competence, autonomy, and relatedness), and the expected behavioral outcomes (more reliable answers; as described for instance in Keusch and Zhang (2015)). Other research questions emerge, such as: how does gamification impact the cognitive load, or would gamification prevent biases (e.g. range insensitivity)?

6. Conclusion

This paper strongly supports our hypothesis that the online elicitation of preferences for MCDA (weights) is feasible and reliable. The presented concept enables reflective thinking and preference construction online and in the absence of an experienced decision analyst. By applying this concept in real population surveys, a wider public audience could be involved in environmental decision-making based on MCDA. More participatory decision-making on complex environmental issues would be possible: the survey concept promotes active citizenship within a participatory democracy paradigm. The survey concept, based on gamification, has been tested and effectively enhances factual learning as well as preference learning, and to some extent the survey experience. The survey concept would benefit from further development according to the outcomes of the present study and as thoroughly discussed in this paper. But the first good news: the proof of concept has been successful.

7. Appendix

The Appendix document presents screenshots of the survey tool (gamified and control versions for the swing and trade-off weight elicitation), the description of the decision problem (objectives, alternatives, and prediction matrix), the ante- and post-questionnaires, the material used for individual assessment in the workshop with ZHdK students, the results of the first exercise of this workshop, the measures and results for hypotheses H1b and H1c, and the reported reasons for the difficulty of the survey results. This document is available free of charge from: xxx.

8. Acknowledgements

We warmly thank Marius Schneider, a research assistant involved in this project. We also thank all participants of the test panel, in particular those who voluntarily provided additional detailed qualitative feedback. Furthermore, we thank René Bauer and Mela Kocher from Zurich University of the Arts (ZHdK) for allowing us to organize the workshops with their students from the game design section, and for interesting discussions. Finally, we thank the team from Takomat GmbH (Köln, Germany: www.takomat-games.com, retrieved on 22.05.2018) for the implementation and maintenance of the prototype during the experiment.

9. Funding

This work is supported by Eawag, the Swiss Federal Institute of Aquatic Science and Technology, as part of the Eawag Postdoctoral Fellowship 2015 (grant number: 5221.00492.009.08 DF 15).

10. Software availability

Name of the software prototype: Eawag survey app. Developers: authors of the paper for the concept, Takomat GmbH (Köln, Germany: www.takomat-games.com, retrieved on 22.05.2018) for the implementation of the prototype. Please contact the first author for further information. Year first available: December 2016. Hardware required: iOS for the prototype version. Software required: download the Eawag survey App from the iTunes store. Availability: open, but the prototype version is protected with a token, please contact the first author. Cost of prototype tester: free of charge. Program language: C#. Program size (App size): 193 Mbyte.

11. References

- Adcock, A., 2012. Cognitive dissonance in the learning process, In: Seel, N.M. (Ed.), Encyclopedia of the Sciences of Learning. Springer: New York Dordrecht Heidelberg London, pp. 2182-2184.
- Amir, O., Levav, J., 2008. Choice construction versus preference construction: The instability of preferences learned in context. *Journal of Marketing Research* 45(2) 145-158.
- Anderson, R.M., Clemen, R., 2013. Toward an improved methodology to construct and reconcile decision analytic preference judgments. *Decision Analysis* 10(2) 121-134.
- Aubert, A.H., Bauer, R., Lienert, J., 2018. A review of water-related serious games to specify use in environmental Multi-Criteria Decision Analysis. *Environmental Modelling & Software* 105 64-78.
- Bailey, P., Pritchard, G., Kernohan, H., 2015. Gamification in market research: Increasing enjoyment, participant engagement and richness of data, but what of data validity? *International Journal of Market Research* 57(1) 17-28.
- Belton, V., Elder, M.D., 1994. Decision support systems: Learning from visual interactive modelling. *Decision Support Systems* 12(4) 355-364.
- Csikszentmihalyi, M., 2000. *Beyond Boredom and Anxiety: Experiencing Flow in Work and Play*. Jossey-Bass, San Francisco, California, USA.
- Deterding, S., 2012. Gamification: designing for motivation. *Interactions* 19 14-17.
- Dillman, D.A., Smyth, J.D., Christian, L.M., 2009. *Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method*, 3rd ed. John Wiley & Sons, Inc., Hoboken, New Jersey, USA & Canada.
- Eisenführ, F., Weber, M., Langer, T., 2010. *Rational Decision Making*. Springer Berlin Heidelberg, Germany, Berlin Heidelberg, Germany.
- Fischer, G.W., 1995. Range sensitivity of attribute weights in multiattribute value models. *Organizational Behavior and Human Decision Processes* 62(3) 252-266.
- Fischhoff, B., 2006. Constructing Preferences from Labile Values, In: Lichtenstein, S., Slovic, P. (Eds.), *The Construction of Preference*. Cambridge University Press New York, pp. 653-667.
- Garrison, D.R., 1997. Self-directed learning: Toward a comprehensive model. *Adult Education Quarterly* 48(1) 18-33.
- Gregory, R., Failing, L., Harstone, M., Long, G., McDaniels, T.L., Ohlson, D., 2012. *Structured Decision Making: A Practical Guide to Environmental Management Choices*. Wiley-Blackwell, West Sussex, UK.

- Gregory, R., Flynn, J., Johnson, S.M., Satterfield, T.A., Slovic, P., Wagner, R., 1997. Decision-pathway surveys: A tool for resource managers. *Land Economics* 73(2) 240-254.
- Gregory, R., Lichtenstein, S., Slovic, P., 2006. Valuing Environmental Resources: A Constructive Approach, In: Lichtenstein, S., Slovic, P. (Eds.), *The Construction of Preference*. Cambridge University Press New York, pp. 1-40.
- Gregory, R., Satterfield, T., Hasell, A., 2016. Using decision pathway surveys to inform climate engineering policy choices. *Proceedings of the National Academy of Sciences* 113(3) 560-565.
- Haag, F., Lienert, J., in preparation. Conceptualizing, eliciting, and evaluating decision objectives for planning of public wastewater infrastructures.
- Haag, F., Lienert, J., Schuwirth, N., Reichert, P., 2018. Identifying non-additive multi-attribute value functions based on uncertain indifference statements. *Omega* Accepted.
- Hämäläinen, R., Kettunen, E., Marttunen, M., Ehtamo, H., 2001. Evaluating a framework for multi-stakeholder decision support in water resources management. *Group Decision and Negotiation* 10(4) 331-353.
- Hämäläinen, R.P., Mustajoki, J., Marttunen, M., 2010. Web-Based Decision Support: Creating a Culture of Applying Multi-Criteria Decision Analysis and Web-Supported Participation in Environmental Decision Making, In: Insua, D.R., French, S. (Eds.), *e-Democracy: a group decision and negotiation perspective*.
- Hamari, J., Koivisto, J., Sarsa, H., 2014. Does Gamification Work? A Literature Review of Empirical Studies On Gamification, 47th Hawaii International Conference on System Sciences: Hawai, USA, pp. 3025-3034.
- Hobbs, B.F., Horn, G.T.F., 1997. Building public confidence in energy planning: a multimethod MCDM approach to demand-side planning at BC gas. *Energy Policy* 25(3) 357-375.
- 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(2) 113-139.
- Irvin, R.A., Stansbury, J., 2004. Citizen participation in decision making: Is it worth the effort? *Public Administration Review* 64(1) 55-65.
- Jorgensen, B.S., Syme, G.J., Smith, L.M., Bishop, B.J., 2004. Random error in willingness to pay measurement: A multiple indicators, latent variable approach to the reliability of contingent values. *Journal of Economic Psychology* 25(1) 41-59.
- Karjalainen, T.P., Rossi, P.M., Ala-aho, P., Eskelinen, R., Reinikainen, K., Kløve, 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(12) 5141-5153.
- Keeney, R.L., 1996. Value-focused thinking: Identifying decision opportunities and creating alternatives. *European Journal of Operational Research* 92(3) 537-549.
- Keeney, R.L., Raiffa, H., 1976. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. Wiley, New York.
- Keusch, F., Zhang, C., 2015. A review of issues in gamified surveys. *Social Science Computer Review* 35(2) 147-166.
- Koivisto, J., Hamari, J., 2014. Demographic differences in perceived benefits from gamification. *Computers in Human Behavior* 35 179-188.
- Lane, D.C., 1992. Modelling as learning: A consultancy methodology for enhancing learning in management teams. *European Journal of Operational Research* 59(1) 64-84.
- Lane, D.C., 1995. On a resurgence of management simulations and games. *The Journal of the Operational Research Society* 46(5) 604-625.
- Langhans, S.D., Reichert, P., Schuwirth, N., 2014. The method matters: A guide for indicator aggregation in ecological assessments. *Ecological Indicators* 45 494-507.

- 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(3) 389-407.
- 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(3) 746-760.
- Marttunen, M., Hämäläinen, R.P., 2008. The decision analysis interview approach in the collaborative management of a large regulated water course. *Environmental Management* 42(6) 1026-1042.
- Massaro, D.W., 2012. Multimodal Learning, In: Seel, N.M. (Ed.), *Encyclopedia of the Sciences of Learning*. Springer: New York Dordrecht Heidelberg London, pp. 2375-2378.
- Merriam, S.B., Caffarella, R.S., Baumgartner, L.M., 2007. *Learning Environments and Learning Concepts, Learning in Adulthood: A Comprehensive Guide*, 3rd ed. John Wiley & Sons, Inc.: San Francisco, CA, USA, pp. 27-52.
- Mezirow, J., 2000. *Learning as Transformation: Critical Perspectives on a Theory in Progress*. Jossey Bass, San Francisco.
- Monks, T., Robinson, S., Kotiadis, K., 2016. Can involving clients in simulation studies help them solve their future problems? A transfer of learning experiment. *European Journal of Operational Research* 249(3) 919-930.
- Montibeller, G., von Winterfeldt, D., 2015. Cognitive and motivational biases in decision and risk analysis. *Risk Analysis* 35(7) 1230-1251.
- Morschheuser, B., Werder, K., Hamari, J., Abe, J., 2017. How to gamify? Development of a method for gamification, 50th Annual Hawaii International Conference on System Sciences (HICSS): Hawaii, USA.
- Mustajoki, J., Hämäläinen, R.P., Marttunen, M., 2004. Participatory multicriteria decision analysis with Web-HIPRE: a case of lake regulation policy. *Environmental Modelling & Software* 19(6) 537-547.
- Niegemann, H.M., Heidig, S., 2012. Multimedia Learning, In: Seel, N.M. (Ed.), *Encyclopedia of the Sciences of Learning*. Springer: New York Dordrecht Heidelberg London, pp. 2372-2375.
- Payne, J.W., Bettman, J.R., Johnson, E.J., 1992. Behavioral decision research: A constructive processing perspective. *Annual Review of Psychology* 43(1) 87-131.
- Pierce, J.L., Kostova, T., Dirks, K.T., 2001. Toward a Theory of Psychological Ownership in Organizations. *The Academy of Management Review* 26(2) 298-310.
- Plass, J.L., Homer, B.D., Kinzer, C.K., 2015. Foundations of game-based learning. *Educational Psychologist* 50(4) 258-283.
- Pöyhönen, M., Hämäläinen, R.P., 2001. On the convergence of multiattribute weighting methods. *European Journal of Operational Research* 129(3) 569-585.
- 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.
- Riabacke, M., Danielson, M., Ekenberg, L., 2012. State-of-the-art prescriptive criteria weight elicitation. *Advances in Decision Sciences* 2012 1-24.
- Rigby, C.S., 2014. *Gamification And Motivation, The Gameful World*. MIT Press: Cambridge, Massachusetts, USA, pp. 113-138.
- Ryan, R., 2017. The motivational pull of video game feedback, rules, and social interaction: Another self-determination theory approach. *Computers in Human Behavior* 73 446-450.

- Ryan, R.M., Rigby, C.S., Przybylski, A., 2006. The motivational pull of video games: A self-determination theory approach. *Motivation and Emotion* 30(4) 344-360.
- Seaborn, K., Fels, D.I., 2015. Gamification in theory and action: A survey. *International Journal of Human-Computer Studies* 74 14-31.
- Sicart, M., 2013. *Beyond Choices : The Design of Ethical Gameplay*. The MIT Press, Cambridge, Mass.
- Slovic, P., 1995. The construction of preference. *American Psychologist* 50(5) 364-371.
- Sweller, J., 2012. Cognitive load theory, In: Seel, N.M. (Ed.), *Encyclopedia of the Sciences of Learning*. Springer: New York Dordrecht Heidelberg London, pp. 2182-2184.
- Tamborini, R., Bowman, N.D., Eden, A., Grizzard, M., Organ, A., 2010. Defining media enjoyment as the satisfaction of intrinsic needs. *Journal of Communication* 60(4) 758-777.
- Thompson, J.P., Howick, S., Belton, V., 2016. Critical learning incidents in system dynamics modelling engagements. *European Journal of Operational Research* 249(3) 945-958.
- Warren, C., McGraw, A.P., Van Boven, L., 2011. Values and preferences: defining preference construction. *Wiley Interdisciplinary Reviews: Cognitive Science* 2(2) 193-205.
- Weber, E.U., Johnson, E.J., 2009. Mindful judgment and decision making. *Annual Review of Psychology* 60(1) 53-85.
- Weber, M., Borchering, K., 1993. Behavioral influences on weight judgments in multiattribute decision making. *European Journal of Operational Research* 67(1) 1-12.