



Decision Support

Can online interfaces enhance learning for public decision-making? Eliciting citizens' preferences for multicriteria decision analysis

Alice H. Aubert^{a,b,*}, Sara Schmid^a, Judit Lienert^a

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

^b Institute of Natural Resource Sciences, Zurich University of Applied Sciences, Wädenswil CH-8820, Switzerland



ARTICLE INFO

Keywords:

Behavioral OR
Learning
Preference elicitation
Online survey
Gamification

ABSTRACT

Innovative online interfaces informing and consulting citizens about their preferences for multicriteria decision analysis (MCDA) could make public decision-making more participatory. We propose a three-faceted learning for decision-making framework and used it to test newly-designed online weight elicitation interfaces. We investigated two features meant to enhance learning: fully-fledged gamification with a narrative, interaction with nonplayer characters, and ambient music, and learning loops (LL) using consistency checks of elicited weights and the challenge to resolve inconsistencies. We operationalized our framework with a novel systematic set of measure instruments providing complementary data types. We designed a 2×2 between-subject experiment with pre- and postquestionnaires. Answers from 769 respondents, representative of the Swiss population in age and gender, indicated that the interfaces successfully raised awareness about wastewater management. Gamification was helpful: respondents performed better in the factual learning test, and unexpected social learning occurred. However, gamification lowered the perception of process understanding. The LL were beneficial: objectively, respondents performed better in the factual learning test. However, respondents perceived the LL as cognitively demanding and their factual learning as lower. Our structured assessment highlighted the need for further research to investigate, for instance, high interpersonal variability and the disparities between tested and perceived learning. Measuring preference construction remains challenging; and social learning should be added to the assessment framework. Applying such structured assessment of learning outcomes to more traditional operational research interventions would provide a baseline for future comparison.

1. Introduction

1.1. General motivation

Making structured decisions, for instance with MCDA (Keeney & Raiffa, 1976) and value-focused thinking (Keeney, 1992), inherently implies learning (e.g. Arvai, Gregory & McDaniels, 2001; Ferretti, Pluchinotta & Tsoukiàs, 2019; Franco & Lord, 2011; Gregory et al., 2012; Linkov et al., 2006; Marttunen & Hämäläinen, 2008). Learning is topical in behavioral OR (Operational Research), and previous studies have focused on evaluating learning in OR interventions (e.g. Monks, Robinson & Kotiadis, 2014, 2016; Thompson, Howick & Belton, 2016). The present work contributes to this endeavor by proposing a comprehensive framework for measuring learning in OR interventions that support decision-making. The framework includes process understanding to the previously proposed framework including factual learning and

preference construction. This enhanced framework is applied to investigate the effects of two newly-designed features meant to enhance learning: gamification and learning loops.

Learning for decision-making is all the more important if participation is broadened. Public engagement is at the core of community OR (e.g. Johnson, Midgley & Chichirau, 2018; Midgley, Johnson & Chichirau, 2018) and evaluating such public interventions has gained attention (e.g. Midgley et al., 2013; White, 2006). Decision analysts beyond community OR have also been concerned with increasing participation (e.g. French & Argyris, 2018; Ríos Insua & French, 2010). Broadening participation allows non-subject-matter experts to engage in decision-making. Ideally, these stakeholders (1) learn facts about the decision topic; (2) construct preferences; and (3) understand the evaluation of alternatives based on objectives that are important to them (value-focused thinking, Keeney, 1992). Knowing that this indeed occurs would increase our confidence in their inputs. This is particularly

* Corresponding author.

E-mail address: aube@zhaw.ch (A.H. Aubert).

<https://doi.org/10.1016/j.ejor.2023.10.031>

Received 24 June 2022; Accepted 26 October 2023

Available online 29 October 2023

0377-2217/© 2023 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

important if the engagement process is happening online because no analyst or moderator can check the reliability of the answers (e.g. Besette, Campbell-Arvai & Arvai, 2016; Philpot, Philpot, Hipel & Johnson, 2022; Vieira, Oliveira & Bana e Costa, 2020). The present work contributes to the effort of increasing citizen participation in decision-making.

We contribute by testing innovative online interfaces to elicit the relative importance given to objectives (weights) for later use in multicriteria decision analysis (MCDA), specifically, multiattribute value theory (MAVT) (Keeney & Raiffa, 1976). We focused on weights, as they are important preference parameters. The present study includes novelties on two features meant to enhance learning: gamification and learning loops. First, we tested a newly designed gamification, based on studies on gamification of learning, and of online surveys (Section 1.2.2). The innovations are the storyline of the tested interfaces, which connects closely to the decision topic at stake, gender-neutral nonplayer characters, and ambient music. The nongamified interface includes the following novelties: a user-friendly interface, and simple instructions that are available at any time. Second, we tested learning loops with consistency checks, in line with decision analysis and learning literature (Section 1.2.3). As innovative contribution, for the first time these learning loops are tested for a real-world decision including a full set of ten objectives.

In addition to the above-mentioned innovations regarding the two features expected to enhance learning (the gamification, and the learning loops), the present study systematically operationalized the measure of the three learning facets. These instruments generate three complementary data types, namely quantitative performance test, quantitative self-assessment, and qualitative self-assessment (Section 2.4). Finally, to investigate the effect of both innovative features on the three learning facets, we carefully designed the experiment, resulting in a 2×2 between-subject experiment with a large and diverse representative population sample.

Below, we present the key related research and our hypotheses (Sections 1.2.1 to 1.2.3; summarized in Section 1.2.4). We continue by presenting the research design and online interfaces in Section 2. We provide the results of our experiment in Section 3. Finally, in Section 4, we summarize our results, reflect on these, and discuss insights for participatory OR and learning through decision-making.

1.2. Summary of key related research

1.2.1. Learning for decision-making

Making structured decisions is often considered as an iterative learning process (e.g. Arvai et al., 2001; Gregory et al., 2012; Linkov et al., 2006; Marttunen & Hämäläinen, 2008). Engaged stakeholders are led into new ways of thinking about their decision problem (Ferretti et al., 2019; Franco & Lord, 2011; Kaplan, 2008) (quotes in Supplementary Material S1). This can be interpreted using the five categories of learning, observed at the individual or group level: change, discovery, explanation, clarification, and creation (Belton & Elder, 1994). Engaged stakeholders can *change* their way of thinking about the problem, *discover* new alternatives and perspectives on the issue, better *explain* the system, and *clarify* or *create* their preferences for themselves and in comparison with preferences of others. Making a structured decision inherently implies learning.

Learning is one of the typically studied aspects within behavioral OR (BOR; Franco, Hämäläinen, Rouwette & Leppänen, 2021). Early work proposes using the value of information to evaluate the expected benefit of learning, in the sense of knowing a decision aspect with more certainty (Clemen, 1991). Information then relates to the probabilities associated with the outcomes of an alternative. New information is valuable if refining the probability changes the decision choice. The value of learning broadens this concept to the whole decision process (McDaniels & Gregory, 2004). Recent BOR publications on learning refer to the theory of action and its learning loop (Monks et al., 2014,

2016). The theory of action is originally used for learning at the level of an organization (Argyris, 1978). Monks et al. (2014) tested the high involvement hypothesis. It suggests that stakeholders who build a model for discrete event simulation learn more than stakeholders who reuse existing models. In a second study, they confirmed that model builders are better at transferring learning from one case to another than model reusers, but model builders are more prone to the overconfidence bias (Monks, Robinson & Kotiadis, 2016). Others refer to constructivist learning theories, in particular transformational learning theory (Mezirow, 2000). They focus on investigating critical learning incidents: those moments of surprise experienced when one's mental model fails to make sense of a situation but a change in the model enhances understanding (Thompson et al., 2016). Our work follows this BOR interest in learning and builds on the constructivist approach.

Given our assumption that preferences are constructed, as opposed to pre-existing and retrieved solely from memory (Lichtenstein & Slovic, 2006), the transformational learning theory (Mezirow, 2000) constitutes an obvious theoretical basis for our work. The transformational learning theory suggests that individuals learn when they actively give meaning to their experience (Mezirow, 2000). This is particularly triggered when individuals face cognitive dissonance (Adcock, 2012) or a disorienting dilemma. Disparities between individuals' existing knowledge or belief and new information lead them to reflect on and adapt their knowledge, mental model, and values. This critical reflection is central to the transformational learning theory. Individuals can critically reflect on the experience itself, termed content reflection; the way they deal with the experience, process reflection; and the long-term established social constructs of the environment in which the experience takes place, premise reflection (Mezirow, 2000).

We suggested conceptualizing an individuals' learning for decision-making as content reflection (Aubert & Lienert, 2019). Seen through the lens of preference construction, we interpret content reflection as twofold. First, individuals need to learn facts about the decision at stake by acquiring information and structuring it so that it becomes useful. This factual learning reinforces or changes pre-existing mental models, or creates one if individuals had no prior internal representation of the decision problem. Second, individuals need to confirm or change values relating to the various aspects of the decision by constructing preferences (Aubert & Lienert, 2019). Then, we proposed a novel interpretation of process reflection for learning for decision-making (Aubert, Esculier & Lienert, 2020, 2022). We considered process reflection as a by-product of preference construction: individuals reflect on the decision-making process and learn how to solve a complex problem, for instance by understanding value-focused thinking. Therefore, we understand individual learning in decision-making as composed of factual learning, preference construction, and process understanding.

We postulate that OR interventions supporting decision-making facilitate these three aspects of learning. We developed new online interfaces to elicit weights for later use in MAVT from affected citizens, who are not subject-matter experts. These interfaces provide state-of-the-art information on objectives and alternatives to enable factual learning. To facilitate preference construction, we follow recommendations to use standard ratio methods such as swing and trade-off for eliciting preferences. These methods emphasize the range of the possible outcomes (Eisenführ, Weber & Langer, 2010; Montibeller & von Winterfeldt, 2015; Riabacke, Danielson & Ekenberg, 2012; von Winterfeldt & Edwards, 1986). Moreover, our interfaces guide the individuals through the elicitation process with instructions that emphasize thinking in terms of objectives (i.e., value-focused thinking) when evaluating possible alternatives. We hypothesize that using the interfaces enables factual learning, preference construction, and process understanding.

1.2.2. Gamification to enhance learning

In addition to testing the occurrence of learning with the developed interfaces, we investigate the effect(s) of two features meant to enhance

learning. The first one is presented in this section, the second in [Section 1.2.3](#). First, we investigate whether gamifying an online survey for preference elicitation of weights for use in MCDA can enhance learning from non-subject-matter experts. Gamification is defined as “the use of game design elements in non-game contexts,” products, and services to motivate desired behaviors ([Deterding, Dixon, Khaled & Nacke, 2011, p.1](#)). Variations of this definition exist, because the concept is relatively new (e.g. [Koivisto & Hamari, 2019](#); [Landers, Auer, Collmus & Armstrong, 2018](#)). Gamification can trigger psychological effects that in turn can lead to effective performance at tasks and persistence in complex actions ([Landers et al., 2018](#); [Ryan & Deci, 2017](#)). Gamification has been tested for multiple purposes, including education and survey research (e.g. [Dichev & Dicheva, 2017](#); [Keusch & Zhang, 2015](#)). We briefly synthesize the state-of-the-art of gamification in these two fields because they are relevant for learning and designing effective online survey interfaces.

Gamification of surveys, sometimes called surveytainment ([Kostyk, Zhou & Hyman, 2019](#)), was initially proposed as a remedy for existing problems such as high dropout rates and low quality of data collected through online surveys. Gamification is expected to maintain the respondents’ interest levels ([Bailey, Pritchard & Kernohan, 2015](#)), which in turn should counter heuristics and other cognitive shortcuts, thereby limiting satisficing behaviors ([Krosnick, 1991](#)) such as random answering and speeding through the survey ([Keusch & Zhang, 2015](#)). However, gamification can distract respondents from the survey, particularly when it is not connected with the survey topic ([Guin, Baker, Mechling & Ruyle, 2012](#)). Other results indicate that gamification leads to a lower overall response rate, but indeed does not introduce biases, and increases the time spent to answer ([Harms, Biegler, Wimmer, Kappel & Grechenig, 2015](#)). Overall, results are equivocal: it is unclear whether gamification of surveys is beneficial. Our work, adapted to the context of decision-making, contributes to answering this open question.

Gamification of learning is also reported to be beneficial in some cases, and detrimental in others. Some review papers in education literature report successes (e.g. [Kasurinen & Knutas, 2018](#); [Subhash & Cudney, 2018](#)). Others have more nuanced views. For instance, in a review of 41 papers on introducing gamification to higher education, only 26% provide empirical evidence for the effectiveness of gamification in improving learning, motivating students, and increasing participation; 64% are inconclusive, and 10% even negative ([Dichev & Dicheva, 2017](#)). Another critical example is the conceptual paper entitled “why gamification fails in education” ([van Roy & Zaman, 2017](#)). These authors propose nine heuristics for designing meaningful gamification with the self-determination theory ([Ryan & Deci, 2017](#)). Examples of these heuristics are to set challenging but manageable goals, and to align gamification with the goal of the activity in question ([van Roy & Zaman, 2017](#)). Gamifying learning is considered promising ([Plass, Homer & Kinzer, 2015](#); [Ryan & Rigby, 2019](#)), but its effectiveness has not yet been conclusively demonstrated.

We developed a new gamified online survey interface for weight elicitation. Hereby, we followed recommendations from the literature and feedback from test users and game designers on our earlier prototypes ([Aubert & Lienert, 2019](#); [Aubert, Lienert, & von Helversen, 2022](#)). While game designers, scientists, and citizens have worked together for modeling and simulation exercises for a long time, such collaboration is innovative for online preference elicitation for MCDA ([Aubert, Bauer & Lienert, 2018](#)). Because gamifying online weight elicitation is so new, we proceeded stepwise by first testing lighter (less expensive) gamification (i.e., not including as many game elements as now), and only with student samples ([Aubert & Lienert, 2019](#)). S4.1 includes a summary table displaying the main difference points with our previous studies. As a result, the new gamification presented in this paper connects the storyline with the survey topic, and provides manageable goals to the respondents, among further new elements presented in [Section 2.2](#) and S4.1 (e.g., gender-neutral nonplayer characters, ambient music). We hypothesized that respondents using the

gamified interface would show improved results for factual learning and preference construction compared to respondents using a control (nongamified) interface. Based on the literature, we could not formulate a clear hypothesis about the effect of gamification on process understanding. Moreover, we tested our hypotheses in a real-world population survey in Switzerland.

1.2.3. Learning loops to enhance learning

Second, we investigated the effect of learning loops for MCDA preference elicitation. Learning loops challenge the respondents to think twice about their preferences. As highlighted earlier when discussing the transformational learning theory ([Mezirow, 2000](#)), cognitive dissonance ([Adcock, 2012](#)) triggers reflection on and adaptation of one’s knowledge, mental model, and values. Facing such instances enhances learning.

In MCDA, recommendations for preference elicitation include asking consistency check questions ([Anderson & Clemen, 2013](#); [Payne, Bettman & Schkade, 2006](#)). In practice, the decision analysts make sure that they elicit preferences that represent the opinion of the stakeholders’ by asking an additional set of questions (e.g. [Hobbs & Meier, 1994](#); [Martin, 2021](#); [Marttunen & Hämäläinen, 2008](#)). For instance, one can use multiple weight elicitation methods and compare the resulting weights. Or, if using swing weight elicitation hierarchically, one can ask about the relative importance of objectives in different branches of the tree, and check if the answer is consistent with the elicited weights. If using pairwise trade-off elicitation, one can ask about a pair of objectives that was not used to calculate the weights, and check if the answer is consistent with the elicited weights. If the stakeholders’ answers are consistent, the preferences elicited are considered reliable. If they are not consistent, the decision analysts guide the stakeholders through the weight elicitation again. The decision analysts can ask the stakeholders to “think harder”. However, in at least one experiment about elicitation of probability distributions from experts, priming experts to think harder was not very effective in making them revise their initial judgments ([Ferretti, Montibeller & von Winterfeldt, 2022](#)). The effect of learning loops created by consistency check questions needs to be investigated.

In education literature, some studies show that tasks requiring high mental effort enhance learning (e.g. [Hamari et al., 2016](#)). Such studies refer to the theories of flow ([Csikszentmihalyi, 1990](#)) and self-determination ([Ryan & Deci, 2017](#)) and suggest that challenges engage the learners if the cognitively demanding tasks match their skills. Thanks to this increased engagement, the challenges lead to greater persistence and better performance at tasks. In the classroom, unchallenging tasks lead to disengagement ([Hamari et al., 2016](#)), which hinders learning. In lifelong learning, persistence in the face of challenges was reported to be facilitated if the learner believes that intelligence is malleable ([Sheffler, Rodriguez, Cheung & Wu, 2022](#)) and perceives internal motivation and self-efficacy ([Merriam & Baumgartner, 2020](#); [Ryan & Deci, 2017](#)). Therefore, if tailored to the learners’ needs, challenges can help create an engaging learning environment.

Some recent online interfaces designed for use without facilitation by a decision analyst create cognitive dissonance to help stakeholders make decisions that are consistent with their own values (e.g. [Bessette et al., 2016](#); [Philpot et al., 2022](#)). They provide feedback on how closely a portfolio or a proposal aligns with their prioritized objectives and values. Resolving the cognitive dissonance and disparities explained in the feedback increase mental effort and stress but also increase process understanding ([Bessette et al., 2016](#)). We proposed a concept of learning loops for online preference elicitation, comparing weights elicited with swing and trade-off methods ([Aubert & Lienert, 2019](#)). However, the experimental design of this previous study did not allow to investigate the effect of this learning loops feature on learning for decision-making.

In the present study, we therefore investigated the effect of learning loops, using newly designed interfaces ([Section 2.2](#)). We implemented two methods for weight elicitation, the swing and trade-off methods. In the online interface, the weights resulting from the two methods are

compared and shown to the respondents. The respondents are then given the opportunity to reconsider their weights in case of inconsistencies (Aubert & Lienert, 2019). In the present study, the instructions are simpler and always accessible, thereby overcoming issues that users reported upon testing previous online weight elicitation interfaces we had developed (Aubert et al., 2020; Aubert, Lienert & von Helversen, 2022). We based our hypotheses on insights from the literature and the idea that learning loops created by consistency checks can provide a “challenge” and cognitive dissonance. We hypothesized that respondents using the interface with learning loops would show better results for factual learning, preference construction, and process understanding than respondents using a control interface (without learning loop). We also hypothesized that learning loops would increase the mental effort.

1.2.4. Measuring the three facets of learning for decision-making

To measure learning, we operationalized a set of instruments (described in Section 2.3). We based our measures on education literature, where the concept of constructive alignment (Biggs, 1996) suggests developing assessments directly connected to each previously defined learning goal. We formulated the following learning goals, based on our proposed three-faceted framework of learning for decision-making: (i) Respondents can remember facts about objectives and alternatives considered in the decision at stake. (ii) Respondents can express consistent preferences with regards to the relative importance of objectives and (iii) are confident that the elicited preferences reflect their actual preferences. (iv) Respondents can differentiate objectives from alternatives and (v) can understand that the relative importance given to objectives is used to derive a ranking of alternatives.

In addition, both BOR literature and education literature call for collecting complementary data types (den Haan & van der Voort, 2018; Franco et al., 2021), i.e., quantitative and qualitative data. Therefore, we designed a set of measurement instruments collecting quantitative (performance test and self-report on Likert scales) and qualitative (text) data for all learning goals. While we could reuse measure instruments from our previous studies for factual learning (e.g., Aubert et al., 2022, 2022), for the present study, we had to revise the measure instruments for preference construction (Aubert et al., 2020, 2022). Moreover, as innovation in this study, we created new instruments for measuring process understanding (i.e. a performance test, and new self-report questions on Likert scales). This comprehensive set of instruments to measure learning for decision-making, derived from a systematic operationalization, is to the best of our knowledge new.

1.2.5. Summary of novelties and hypotheses

This study presents several innovations. First and second novelties relate to the features meant to enhance learning. Specifically, the tested interfaces – gamified and with learning loops – are new, and their respective design builds up on lessons learnt from previous attempts. Additionally, the present experimental design, with a large representative population sample, allows to investigate the effect of the learning loops feature specifically. Finally, we propose a comprehensive novel set of instruments to measure learning for decision-making. Table 1 summarizes the hypotheses formulated in the introduction, following our proposed conceptualization of individual learning for decision-making. It is worth mentioning that based on the literature, we could not formulate clear hypothesis about the effect of gamification on process understanding, nor about the interaction of both features meant to enhance learning.

2. Methods

2.1. Decision case: wastewater infrastructure

Our experiment addresses a relevant decision about wastewater infrastructure planning. Today's centralized wastewater system has

Table 1
Summary of hypotheses. LL: learning loop.

Facets of learning	Hypotheses
(1) Factual learning	(1a) Answering specifically designed online survey interfaces for decision-making enables factual learning.
	(1b) Using a gamified online survey interface improves factual learning compared to using a nongamified interface.
	(1c) Using an online survey interface with LL improves factual learning compared to using an interface without LL.
(2) Preference construction	(2a) Answering specifically designed online survey interfaces for decision-making enables preference construction.
	(2b) Using a gamified online survey interface improves preference construction compared to using a nongamified interface.
	(2c) Using an online survey interface with LL improves preference construction compared to using an interface without LL.
(3) Process understanding	(3a) Answering specifically designed online survey interfaces for decision-making enables process understanding.
	(3b) <i>Based on the literature, we could not formulate a clear hypothesis about the effect of gamification on process understanding.</i>
	(3c) Using an online survey interface with LL improves process understanding compared to using an interface without LL.
(4) Extraneous cognitive load (ECL)	(4) LL increase the mental effort compared to no LL.

been highly successful, but also has drawbacks regarding sustainability and costs. Increasingly, decentralized wastewater systems are being developed and discussed (Hoffmann et al., 2020; Larsen, Hoffmann, Lüthi, Truffer & Maurer, 2016). These are no longer connected to sewers; instead wastewater is treated on-site in people's homes. Wastewater infrastructure needs renewal in many OECD countries, and especially in rural municipalities, this offers a window of opportunity for transitioning to decentralized systems. For our experiment, we used information from a decision about wastewater infrastructure planning in two Swiss rural municipalities (Beutler & Lienert, 2019). Deciding about future wastewater management usually affects the inhabitants via wastewater fees. Additionally, decentralized wastewater infrastructure can directly affect people in their homes. The decision in this experiment consisted of choosing between six alternatives, based on preferences regarding ten objectives. Based on information collected in the Swiss cases (Beutler & Lienert, 2019) and expertise from earlier wastewater projects (Haag, Zürcher & Lienert, 2019; Lienert, Duygan & Zheng, 2016), we considered four higher-level objectives, which we specified with two to three lower-level objectives (Fig. 1). The six alternatives were: rehabilitation of the local wastewater treatment plant (WWTP), connection to the regional WWTP, decentralized package plants, decentralized package plants with urine separation, decentralized package plants with urine and feces separation, and construction of a new local WWTP with urine separation (see S2 and S3 for a detailed description of the objectives and alternatives).

2.2. The experiment: procedure and treatments

We tested if the three aspects of learning occurred while using the specifically designed interface to elicit weights, and evaluated the gamification and learning loops effects as hypothesized in the introduction (Table 1). To test this, we designed a 2 × 2 between-subject experiment. The two varying factors defining our four treatments were nongamified (control) vs. gamified interface, and without (control) vs. with learning loops (Fig. 2). Note that we could not exclude potential interactions between the treatments. We briefly introduced the survey topic (wastewater management) and explained the survey procedure.











1. Low pollution of water bodies
 - 1.1 High removal of nitrogen compounds 
 - 1.2 High removal of micropollutants 
2. High use of natural resources
 - 2.1 High recovery of phosphorus 
 - 2.2 Low use of water 
 - 2.3 Low net energy consumption 
3. High social acceptance (for users)
 - 3.1 High health protection 
 - 3.2 High attractiveness of toilets 
 - 3.3 Low time demand for end-users 
4. Low costs for this and future generations
 - 4.1 Low annual operating and maintenance costs 
 - 4.2 High flexibility (intergenerational equity) 

Fig. 1. Hierarchy of objectives.

Thereafter, respondents completed a knowledge test about wastewater management in Swiss rural contexts. One of the four treatments followed, in which we informed all respondents about the ten objectives and six possible alternatives and elicited how respondents weighed the objectives. In the two treatments with learning loops, respondents also ranked the alternatives from most to least preferred after reading about the alternatives and before weight elicitation. Thereafter, we directed respondents to a post-treatment questionnaire that included a repetition of the knowledge test, performance tests, self-reported assessments, open textbox questions, and thanks (Fig. 2).

2.2.1. Nongamified (control) vs. gamified treatments

The nongamified treatment (Fig. 3) was an interface specifically designed for weight elicitation (Aubert & Masson, 2021). It included improvements of two previously tested prototypes (Aubert et al., 2020, 2022). It described the objectives and alternatives and implemented swing weight elicitation following textbook recommendations (Anderson & Clemen, 2013; Eisenführ et al., 2010; Payne et al., 2006). Novel features to overcome earlier problems included instructions with a graphic example, which remained accessible at all times. This clarified instructions, and avoided repetitive text. Moreover, the new user interface was simpler than the previous interfaces and included a progress bar. It also provided warning messages. For instance, in swing weight elicitation, a message popped up if the respondent intended to rate a lower-ranked alternative higher than a higher-ranked one (S4.1, screenshots in S4.2). Finally, although of lesser importance for the

experiment, the new interface included an administration back-end, which significantly eased creating and modifying surveys as no coding knowledge is required.

The gamified treatment (Fig. 3) consisted of an interface that included game elements (also see screenshots in S4.3). A narrative provided a goal, and interactions occurred with gender-neutral non-player characters. The innovation of using gender-neutral characters was based on feedback from earlier prototype testing and game designers (Aubert & Lienert, 2019). The novel narrative was strongly anchored in the wastewater decision topic. The survey respondents were candidates in the election to become the mayor of New Waterton. In the neighboring town, wastewater management problems had occurred. New Waterton’s citizens wanted to avoid these issues, and the wastewater topic became decisive for their vote. To increase their chances of being elected, the candidates (i.e. respondents) needed to communicate consistently about their preferences. As another new feature, the respondents chose an avatar and provided a name for the exchanges with the nonplayer characters. Each nonplayer character represented a citizen who advocated a single objective. There were additional nonplayer characters who gave feedback and guidance, and challenged the respondents to do better when relevant. To ensure gender neutrality, another novel aspect, characters resembled animals. The interface blended weight elicitation into the narrative, thereby taking some minor liberties for the sake of simplicity regarding textbook recommendations for weight elicitation (S4.4). The interface included original artwork and ambient music that corresponded with the narrative, an innovation not included in our earlier prototypes. It provided more choices to the respondents than the nongamified treatment. For instance, respondents decided on the order of tasks within a chapter. It rewarded the respondents (e.g. with feedback from nonplayer characters, or by being elected). As the nongamified version, it included a progress bar. The informative texts about the ten objectives and the six alternatives were strictly the same as in the nongamified treatment. The objectives’ icons and the graphical description of alternatives were also the same.

After a competitive call, the company Entrée de Jeux (<https://www.entree-de-jeux.ch/>, retrieved on 14.6.2022) conceptualized the gamification, and the company Opinion Games (<https://www.opiniongames.ch/>, retrieved on 14.6.2022) developed it. The company Youmi (<https://www.youmi-lausanne.ch/>, retrieved on 14.6.2022) developed the nongamified interface. Both developments closely involved the first author. As a small aside, we wish to point out that this collaboration between science and two game companies was complex, time-consuming, and expensive.

2.2.2. Without (control) vs. with learning loops

In the treatment with learning loops, we elicited preferences with two methods: swing and trade-off in series (for method description, see e.g. Eisenführ et al., 2010). After elicitation, the two sets of elicited weight preferences were presented to the respondents. Comparing the results originating from the two methods allowed respondents to reflect and think twice about their weight preferences. This comparison step mimics consistency check questions by a facilitator in direct interactions (e.g. Hobbs & Meier, 1994; Payne et al., 2006). After comparing the two sets of elicited weights, the respondents could decide to repeat the elicitation and change their answers, which created a learning loop. We calculated a ranking of alternatives using the additive aggregation

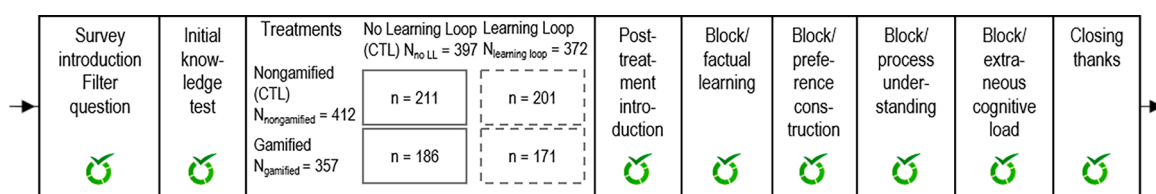


Fig. 2. Experiment flow and sample sizes (after data cleaning). CTL: control treatment; LL: learning loop.

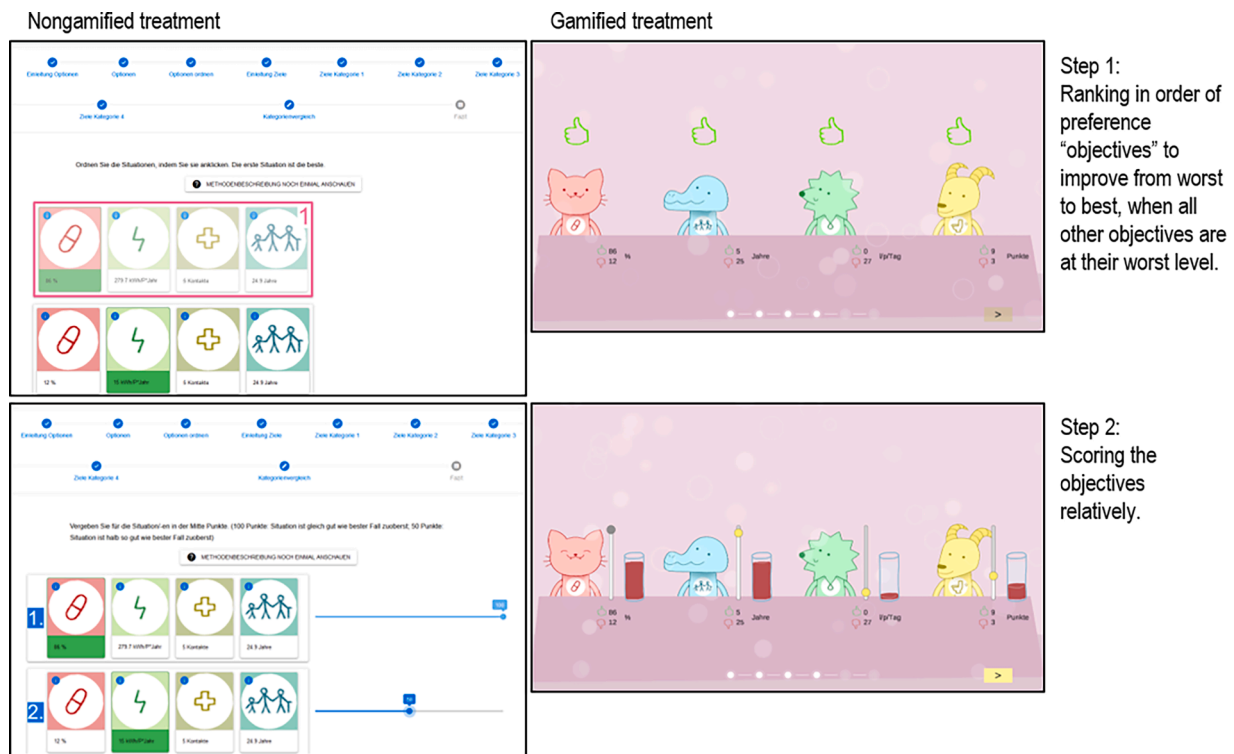


Fig. 3. Examples of screenshots of the interfaces in the nongamified (left) and gamified treatment (right). It displays the swing weight elicitation.

model of multiattribute value theory (MAVT), linear marginal value functions, and the elicited weights (Keeney & Raiffa, 1976). This ranking was compared with holistic ranking of alternatives from most to least preferred provided by the respondents before weight elicitation but after reading about the alternatives. The respondents also received a graphical representation of the prediction matrix that showed how well each alternative performed on each objective. The concept of these two embedded learning loops has been pretested and is described in detail in Aubert and Lienert (2019). The treatment without learning loop consisted of an initial holistic ranking of alternatives, swing weight elicitation, and a final ranking of alternatives based on the elicited weights and the additive MAVT model without feedback.

2.3. Sample definition and respondent recruitment

A market research company (Intervista; <https://www.intervista.ch/>, retrieved on 14.6.2022) invited the respondents by email. In the email, they mentioned the unusual length of the survey, at least 45 min, and that it was only possible to complete it from a desktop computer or laptop. Respondents received a link to the opening part of the experiment. Intervista made sure to invite each respondent only to a single treatment, and assigned them to one of the four treatments. We mandated the company to invite respondents from the German-speaking part of Switzerland, following age and gender quota to represent the Swiss population in each treatment (also see Section 3.1 and S5). Upon completion, respondents received a token compensation according to the company's point system. Because we targeted non-subject-matter experts, the survey started with a filter question asking how much respondents knew about wastewater management. If they knew *rather a lot* or *a lot*, they could not proceed to the survey. We excluded the knowledgeable respondents, because we assumed that they would not learn much about wastewater management. However, this could be the topic of a follow-up experiment. In the present study, only respondents knowing *nothing at all* or *a little* about wastewater management could answer because our survey was explicitly designed for informing and gathering preferences of laypeople.

We determined the desired sample size from previous studies (Ryan, 2017) and a priori statistical power analysis (Faul, Erdfelder, Buchner & Lang, 2009). We aimed at 200 respondents per treatment. Each subsample was comparable in age, gender, and education distributions (Section 3.1, S6 and S7). We collected data for the nongamified treatments between March and April 2021 and for the gamified treatments between October and November 2021.

2.4. Variables and measures

To measure the three aspects of learning for decision-making, we collected complementary data types (den Haan & van der Voort, 2018; Franco et al., 2021): quantitative, as performance tests; quantitative, as self-reported assessments on 5-point Likert scales; and qualitative, as optional open textbox questions (Table 2, S8). The extraneous cognitive load was measured on 5-point Likert scales (Table 2, S8). The measures were found in the literature and/or improved from our previous experiments (Aubert & Lienert, 2019; Aubert et al., 2022; Danilenko, 2010; Leppink, Paas, Van der Vleuten, Van Gog & Van Merriënboer, 2013, 2014; Lienert et al., 2016). Measures for process understanding were newly created.

2.5. Data analysis

The statistical analyses for the quantitative data were performed with R project for statistical computing (R Development Core Team, 2020). We looked at central tendencies and distributions. We investigated whether the treatments explained the measures using linear regression analyses for the performance tests (function *lm*; $performance = f(gamified, learning_loop, gamified * learning_loop)$) and beta regressions for the Likert scales measures, because they are constrained (function *betareg*; $self-reported = f(gamified, learning_loop, gamified * learning_loop)$). Beta regressions required normalizing the 1 to 5 scales. We controlled for outlying observations with studentized residuals and Cook's distance, absence of multicollinearity (i.e. absence of linear correlation between the explanatory variables) with the variance inflation factor

Table 2
Measures used to assess the three aspects of learning.

Learning goals	Type	Measures
1. Factual learning Respondents can remember facts about objectives and alternatives considered in the decision at stake. (S8.1)	Quantitative/ performance test	- Ten multiple-choice questions about the facts concerning the objectives to consider when deciding about wastewater management (one per objective). For each question, four possible answers were available. Respondents were asked to choose the correct statements among the four propositions. Order of questions and choices for each question was randomized. In the initial test, there was a fifth option: 'I do not know'. Rating system: 1 point if responses to all four choices were correct (which meant the choice was correctly checked or correctly not checked); 0.5 points if three choices were correct; 0 points otherwise (also 0 points if 'I do not know' was chosen). Knowledge score for objectives varied from 0 to 10. - Six multiple-choice questions about the alternatives to choose from in our decision case, one per alternative. As for objectives, order of questions and choices for each question were randomized. Same rating system as for the objectives. Knowledge score for alternatives varied from 0 to 6.
	Quantitative/ self-reported	- Self-reported confidence in the answers provided for the knowledge test (from 1 <i>Not at all confident. I am 1 percent confident.</i> to 5 <i>Very confident. I am 99 percent confident.</i>) - Three self-reporting 5-point Likert scale questions on factual learning (from 1 <i>no learning at all</i> to 5 <i>very much learning</i>). Self-reported factual learning was calculated as the mean of the three questions and varied from 1 to 5.
	Qualitative	Optional textbox question asking <i>Please elaborate on the facts you have learnt about wastewater management</i> . The answers were coded right or wrong per objective and alternative. Statements indicating a preference change were coded as well.
2. Preference construction Respondents can express consistent preferences with regards to the relative importance of objectives and are confident that the elicited preferences reflect their actual preferences. (S8.2)	Quantitative/ performance test	- Two multiple-choice questions about the respondent's preferences on objectives (one for the lower level of the objectives hierarchy, one for the upper level). For each question, four possible answers were available. Respondents were asked to choose the correct statements among the four propositions. Order of the choices for each question was randomized. Same rating system as for the objectives. Preference construction score varied from 0 to 2. - Percentage of respondents with weight patterns: (i) indifferent between all objectives (equal), (ii) one objective receiving a weight over 0.95 (single).
	Quantitative/ self-reported	Three self-reporting 5-point Likert scale questions on preference construction (from 1 <i>no preference construction at all</i> to 5 <i>very much preference construction</i>). Self-reported preference construction was calculated as the mean of the three questions and varied from 1 to 5.
	Qualitative	Optional textbox question asking <i>Please elaborate what exactly you have learnt about your preferences</i> . The answers were coded per objective, and alternative. Statements indicating factual learning were coded as well.
3. Process understanding Respondents can differentiate objectives from alternatives and can understand that the relative importance given to objectives is used to derive a ranking of alternatives. (S8.3)	Quantitative/ performance test	Four multiple-choice questions about value-focused thinking, the meaning of objectives and alternatives, and the task performed. For each question, four possible answers were available. Respondents were asked to choose the correct statements among the four propositions. Order of the choices for each question was randomized. Same rating system as for the objectives. Process understanding score varied from 0 to 4.
	Quantitative/ self-reported	Four self-reporting 5-point Likert scale questions on process understanding (from 1 <i>no understanding at all</i> to 5 <i>very much understanding</i>). Self-reported process understanding was calculated as the mean of the four questions and varied from 1 to 5.
	Qualitative	Two optional textbox questions: <i>Please elaborate what exactly you have learnt about the way to tackle complex decision problems</i> and <i>Will you use this method to tackle a complex decision in a different context? Describe shortly how you would do it</i> . The answers were coded per objective and alternative, and as being correct or wrong. Statements indicating a preference change were coded as well.
4. Extraneous cognitive load (ECL) Is there an effect of the gamification on ECL? Is there an effect of learning loops on ECL?	Quantitative/ self-reported	Four self-reporting 5-point Likert scale questions on the mental effort required for process understanding (from 1 <i>no ECL at all</i> to 5 <i>very much ECL</i>). Self-reported ECL was calculated as the mean of the four questions and varied from 1 to 5. (S8.4)

(VIF) and tolerance, and independent errors with absence of autocorrelation and the Durbin-Watson test. The checks for assumptions were acceptable for all models.

The qualitative data were coded in an iterative process. We retained the qualitative answers of 833 respondents, assuming that if they took the time to write something, their input could be useful even if they showed speeding behavior in the previous questions and had been excluded from the main analysis (Section 3.1). The answers to the four questions (Table 2) were redistributed according to the three aspects of learning for decision-making. The coding scheme and data are freely available in the data package (<https://doi.org/10.25678/0008WS>).

After coding, we produced contingency tables for the occurrence of the codes and corrected the counts by the number of respondents in each treatment. We report hereafter only on those codes that differ by a factor of 2 between treatments.

3. Results

3.1. Respondents and pre-analyses

In total, 2094 respondents started the survey (Table 3). We kept only the complete responses. The length of the survey induced high dropout:

median time varied from 42 min for the nongamified treatment without learning loop to 65 min for the gamified treatment with learning loop. Unfortunately, during data collection for the gamified treatment, the server stopped working, which prevented access. This contributed to a higher loss of respondents for the gamified treatment. We removed 51 respondents with satisfying behaviors such as speeding (less than half the median time) and straightlining (e.g. respondents choosing the same value on Likert scales for questions where it does not make sense). We excluded five respondents who wrote dubious comments, such as “One always eats kebab with cocktail sauce.” The final sample size was 769 respondents (Fig. 2, Table 3); 54% identified as female, 46% as male; age varied between 18 and 84 (the average age of the sample was 49); most (45%) had a university degree (full description in tables in S6). In the analyzed sample, there was no statistical difference between the four treatment groups regarding gender, age and education (S6). Additionally, there was a balanced distribution of these covariates across the treatments (S7); we thus did not use the socio-demographic variables in the analyses.

3.2. Factual learning

3.2.1. Performance test for learning about objectives

Overall, the final scores for factual learning were significantly higher than the initial scores by 1.5 points of 10 total points ($t(768) = 23.806, p < .001, d_{Cohen} = 0.81$). Therefore, the respondents’ delta knowledge scores ($\Delta KS = KS_{final} - KS_{initial}$) were positive on average (Table in S9.11). For our samples, the KS increased by 1.22 to 1.88 points. This indicates that only limited learning about the objectives occurred. The multiple linear regression analysis showed that gamification ($\beta = -0.55, p = .002$) and learning loop ($\beta = -0.39, p = .03$) had an effect on factual learning about objectives ($F(3, 765) = 5.33, p < .001$): the nongamified treatment had a lower delta score than the gamified treatment by 0.55 points of 10 total points; the sample without learning loop had a lower delta score than the sample with learning loops by 0.39 points. However, the interpersonal variability was high: only a very small proportion of the variance was explained ($R^2 = 1.7\%$) (for reasons of space, the full regression table is given in Tab.S912). Finally, fewer respondents had negative ΔKS or null ΔKS in the gamified treatments than in the nongamified ones (Table 4). In other words, in the nongamified treatments, more respondents had lower knowledge scores at the end of the survey than at the beginning. The binomial logistic regression ($\chi^2(3) = 13.245, p = .004$) showed that gamification was a significant predictor of negative ΔKS ($\beta = 0.65, z = 2.51, p = .012$); while the learning loop and the interaction (gamified*learning loop) terms were not significant predictors (Tab.S913)

3.2.2. Performance test for learning about alternatives

Overall, the final scores were significantly higher than the initial by 0.6 points of total 6 points possible ($t(768) = 13.289, p < .001, d_{Cohen} = 0.59$). Thus, the respondents’ delta knowledge scores for alternatives ($\Delta KS_{alt} = KS_{alt final} - KS_{alt initial}$) were positive but low (Tab.S914). For our samples, KS_{alt} increased by 0.41 to 0.71 points. Learning about the alternatives occurred, but was limited. The multiple linear regression analysis showed no statistically significant effect of the treatments on factual learning about alternatives ($F(3, 765) = 2.37, p = .07$) (Tab. S915). The proportions of negative ΔKS_{alt} or null ΔKS_{alt} did not differ

Table 3

Statistics of respondents. Start: the number of respondents who started the survey; Complete: the number of respondents who answered until the end (% of those who started in parentheses); Sample: the number of respondents after data cleaning (removing speeding, straightlining, dubious comments) (% of those who started in parentheses),%Lost: the proportion of respondents in the sample relative to those who started. LL; learning loop. nogam: nongamified. gam: gamified.

	Start		Complete		Sample		% Lost	
	noLL	LL	noLL	LL	noLL	LL	noLL	LL
nogam	375	459	242 (64.5%)	236 (51.4%)	211 (56.3%)	201 (43.8%)	43.7	56.2
gam	411	849	206 (50.1%)	205 (24.1%)	186 (45.3%)	171 (20.1%)	54.7	79.9

much between the treatments (Table 4). However, the logistic regression ($\chi^2(3) = 9.00, p = .029$) showed that the interaction term (gamified*learning loop) was a significant predictor of the negative ΔKS_{alt} ($\beta = 0.72, z = 2.43, p = .015$): respondents with nongamified treatment without learning loop had more often negative ΔKS_{alt} , meaning that their KS_{alt} at the end were lower than at the start (Tab.S916).

3.2.3. Self-reported factual learning

Overall, the respondents’ self-reported factual learning was slightly positive (3: fairly, 4: rather better/increased on 5-point Likert scale; Table 5). Respondents with learning loops had the lowest self-reported learning. The beta regression analysis showed that the learning loop had an effect on self-reported factual learning ($\beta = 0.18, p = .002$). To a lesser extent, the interaction between gamified and learning loop also had an effect ($\beta = -0.13, p = .098$). However, again, the interpersonal variability was high: little of the variance was explained for the sample without learning loop that had slightly higher self-reported learning ($R^2 = 2.0\%$) (Tab.S917).

Overall, the respondents’ confidence in their answers to the knowledge test slightly increased at the end of the survey compared to the beginning (by 0.49 on average, $t(768) = 17.229, p < .001, d_{Cohen} = 0.58$; Table 6). The confidence increased more for respondents without learning loop than those receiving the learning loops and for respondents with gamification than those receiving the nongamified treatment on average (Table 6). The beta regression analysis showed that the learning loop was a predictor of self-reported confidence in the answers to the knowledge test ($\beta = 0.06, p = .073$). However, the interpersonal variability was high: little of the variance was explained ($R^2 = 1.6\%$) (Tab.S918).

3.2.4. Qualitative data

Ten percent more of the respondents receiving the gamified treatment (55.3%) left comments that were coded as correct than those receiving the nongamified treatment (44.9%) ($\chi^2(1) = 9.060, p = .003$). More than twice as many respondents receiving the gamified treatment (8.1%) acknowledged the complexity of the wastewater system than the respondents receiving the nongamified treatment (3.6%) ($\chi^2(1) = 7.808, p = .006$), for instance “that wastewater management is a complex business: it requires lots of expert knowledge to be able to make a good decision,” “Complex system with many dependencies and conflicting objectives,” “it is a much bigger and more complex issue than I first thought.” More than twice as many respondents receiving the nongamified treatment (3.1%) wrote a negative comment suggesting that the information was unclear or that there was too much information than the respondents receiving the gamified treatment (1.0%) ($\chi^2(1) = 4.262, p = .054$). For instance: “the explanations were complicated... at least for me as a beginner,” “too technical and too difficult,” “difficult to understand for laypeople”. Answers from treatments with and without the learning loop did not differ much.

3.3. Preference construction

3.3.1. Performance test for preference construction

Overall, the participants’ preference construction score was 0.84 (ranging from 0 to 2; Tab.S921). The lowest score was obtained for the respondents with the nongamified treatment and no learning loop (but

Table 4

Number of respondents (and percentage) with a negative (lower final knowledge score compared to initial) and null Δ KS (no learning) for objectives and alternatives. Δ KS = final KS – initial KS; Δ KS_{obj} can vary between –10 and 10. Δ KS_{alt} can vary between –6 and 6. obj: objectives. alt: alternatives. noLL: no learning loop. LL: learning loop. nogam: nongamified. gam: gamified. *Statistically significant.

		noLL		LL		Total	
		Δ KS < 0	Δ KS = 0	Δ KS < 0	Δ KS = 0	Δ KS < 0	Δ KS = 0
Obj.	nogam	41 (19.43%)	26 (12.32%)	35 (17.41%)	20 (9.95%)	76 (18.45%)	46 (11.17%)
	gam	27 (14.52%)	16 (8.60%)	15 (8.77%)*	13 (7.60%)	42 (11.76%)	29 (8.12%)
	total	68 (17.13%)	42 (10.58%)	50 (13.44%)	33 (8.87%)	118 (15.34%)	75 (9.75%)
Alt.	nogam	56 (26.54%)	47 (22.27%)*	50 (24.88%)	23 (11.44%)	106 (25.73%)	70 (16.99%)
	gam	41 (22.04%)	26 (13.98%)	42 (24.56%)	28 (16.37%)	83 (23.25%)	54 (15.13%)
	total	97 (24.43%)	73 (18.39%)	92 (24.73%)	51 (13.71%)	189 (24.58%)	124 (16.12%)

Table 5

Self-reported factual learning. Average from three items (Cronbach’s alpha* = 0.81); min = 1; max = 5. SD = standard deviation. Mdn = Median. noLL: no learning loop. LL: learning loop. nogam: nongamified. gam: gamified.

	noLL				LL				total			
	Min	Max	Mdn	Mean(SD)	Min	Max	Mdn	Mean(SD)	Min	Max	Mdn	Mean(SD)
nogam	1	5	3	3.20 (0.70)	1.3	5	3	3.14 (0.65)	1	5	3	3.17 (0.67)
gam	1	5	3.33	3.40 (0.68)	1	4.67	3.33	3.18 (0.64)	1	5	3.33	3.29 (0.67)
total	1	5	3	3.29 (0.70)	1	5	3	3.16 (0.64)	1	5	3.33	3.23 (0.68)

* Cronbach’s alpha is a measure of how closely related the items of a scale are. If it is close to 1, the items are reliably related and can be averaged.

Table 6

Change in confidence in the answers from the knowledge tests. Difference from confidence in the answers in post-test–pretest; Varied from min = –4 to max = 4. SD = standard deviation. Mdn = Median. noLL: no learning loop. LL: learning loop. nogam: nongamified. gam: gamified.

	noLL				LL				total			
	Min	Max	Mdn	Mean(SD)	Min	Max	Mdn	Mean(SD)	Min	Max	Mdn	Mean(SD)
nogam	–2	2	0	.44 (0.78)	–4	2	0	.38 (0.83)	–4	2	0	.41 (0.81)
gam	–1	2	1	.65 (0.72)	–1	3	0	.49 (0.75)	–1	3	1	.57 (0.74)
total	–2	2	1	.54 (0.76)	–4	3	0	.43 (0.80)	–4	3	0	.49 (0.78)

the differences were not statistically significant, Tab.S922). We checked for suspicious patterns in weights: equal weights, or nearly or all weight assigned to one objective (Table 7). A general linear regression ($\chi^2(3) = 51.726, p < .001$) showed that neither gamification, learning loop, nor their interaction were good predictors of the occurrence of patterns (Tab.S923). However, only two patterns occurred without learning loop, against 40 with learning loops.

3.3.2. Self-reported preference construction

Overall, the respondents’ self-reported factual learning was slightly positive (3: fairly, 4: rather better/increased on a 5-point Likert scale; Table 8). The pre-analysis for reliability between items resulted in a dubious Cronbach’s alpha (below 0.7); therefore, we did not aggregate them in one construct. Respondents with learning loops generally had lower self-reported preference construction than those without, except for item 2 gamified (Table 8). The beta regression models to explain items 1 and 2, both about the perceived occurrence of preference construction, showed no significant effect of the treatment (Tab.S924, S925). The beta regression model to explain item 3, about confidence in the elicited preferences, showed some effect of learning loop ($\beta = 0.14, p = .070$): with the learning loops, the confidence was lower (Tab.S926). Again, little of the variance was explained ($R^2_{item3} = 0.92\%$). Results about self-reported preference construction could not support nor contradict our assumptions.

3.3.3. Qualitative data for preference construction

The proportion of respondents with negative comments about preference construction was higher (22.3%) for those receiving the gamified treatment than the nongamified treatment (14.3%) ($\chi^2(1) = 9.092, p = .003$). More than twice as many respondents receiving the gamified treatment (6.5%) wrote a comment suggesting that they faced

difficulties in trading off objectives than those receiving the nongamified treatment (3.1%) ($\chi^2(1) = 5.265, p = .031$). This is particularly true for those with learning loops: 8.6% for gamified vs. 3.7% for nongamified ($\chi^2(1) = 4.214, p = .056, n.s.$). For instance, they wrote how important two objectives were, despite the fact that these objectives cannot be achieved at the same time: e.g. “Saving energy and water does not go both ways,” “that it is difficult to reconcile environmental protection with finances and attractiveness for the population,” “It’s like having to choose between eating or drinking.” Finally, more than twice as many respondents receiving the gamified treatment (7.3%) wrote a comment suggesting population acceptance issues and conflicting opinions than those receiving the nongamified treatment (1.5%) ($\chi^2(1) = 16.773, p < .001$), for instance “Time cost for the end-user should be close to 0, else the system won’t be accepted,” “I did not think that wastewater management should consider the attractiveness of toilets to increase acceptance,” “User acceptance is very important” (more

Table 7

Number of respondents and percentage with weight patterns. Equal: $W_{max} - W_{min} < 0.05$ (indifference between objectives). Single: $W_{max} > 0.95$ (a single objective receives nearly or all the weight). noLL: no learning loop. LL: learning loop. nogam: nongamified. gam: gamified.

	noLL		LL		total	
	Equal	Single	Equal	Single	Equal	Single
nogam	2 (0.95%)	0 (0%)	14 (6.97%)	3 (1.49%)	16 (3.88%)	3 (0.73%)
gam	0 (0%)	0 (0%)	20 (11.70%)	3 (1.75%)	20 (5.60%)	3 (0.84%)
total	2 (0.50%)	0 (0%)	34 (9.14%)	6 (1.61%)	36 (4.68%)	6 (0.78%)

Table 8

Self-reported preference construction. This is reported at the item level without aggregation because of dubious Cronbach’s alpha = 0.63; min = 1; max = 5. SD = standard deviation. Mdn = Median. noLL: no learning loop. LL: learning loop. nogam: nongamified. gam: gamified.

	noLL				LL				total			
	Min	Max	Mdn	Mean(SD)	Min	Max	Mdn	Mean(SD)	Min	Max	Mdn	Mean(SD)
nogam												
Item 1	1	5	3	3.13 (0.74)	1	5	3	3.09 (0.76)	1	5	3	3.11 (0.75)
Item 2	1	5	4	4.00 (0.88)	1	5	4	3.95 (0.78)	1	5	4	3.97 (0.83)
Item 3	1	5	3	3.26 (0.95)	1	5	3	3.13 (0.82)	1	5	3	3.20 (0.89)
gam												
Item 1	1	5	3	3.16 (0.69)	1	5	3	3.12 (0.76)	1	5	3	3.14 (0.72)
Item 2	2	5	4	4.01 (0.75)	2	5	4	4.01 (0.75)	2	5	4	4.01 (0.75)
Item 3	1	5	3	3.32 (0.85)	1	5	3	3.15 (0.91)	1	5	3	3.24 (0.88)
total												
Item 1	1	5	3	3.14 (0.72)	1	5	3	3.11 (0.76)	1	5	3	3.13 (0.74)
Item 2	1	5	4	4.01 (0.82)	1	5	4	3.98 (0.76)	1	5	4	3.99 (0.79)
Item 3	1	5	3	3.29 (0.90)	1	5	3	3.14 (0.86)	1	5	3	3.22 (0.89)

Item 1: How much did you learn about your preferences (i.e., by realizing, that some objectives are more important than others)? Item 2: In the preference elicitation, did you answer in a way that reflects your personal preference? Item 3: How confident are you about the preferences that you expressed during the preference elicitation?.

quotations in S927).

However, it should be noted that a small proportion of respondents used the term “preference” imprecisely (varying between 0.9 and 4.8%), for instance “When different preferences have the same priority,” “I became aware of the weighting of my preferences in comparison to each other,” “I rate individual preferences higher than others”. Eight respondents, seven from the nongamified and one from the gamified treatment, expressed doubts about the reliability of the elicited preferences, for instance “Would be interested to know how many results can be effectively used,” “it will be difficult to collect reliable and valid results,” “the final results are therefore rather falsified”. Finally, 5.1% of the respondents receiving learning loops reported no preference construction compared to 0.5% respondents receiving the treatment without learning loop ($\chi^2(1) = 15.869, p < .001$), for instance “nothing new learnt,” “little to nothing, rather confusion,” “actually, I have not yet reflected about preferences in wastewater management!”. More results are presented in S927.

3.4. Process understanding

3.4.1. Performance test for process understanding

Overall, the respondents’ process understanding score was 2.06 (ranging from 0 to 4; Tab.S931). The highest score was obtained by the respondents with the gamified treatment with learning loops in the multiple linear regression (by 0.17 points, $p = .073$) ($R^2 = 0.2\%$) (Tab. S932).

3.4.2. Self-reported process understanding

Overall, the respondents’ self-reported process understanding was slightly positive (3: fairly, 4: rather better/increased on a 5-point Likert scale; Tab.S933). The multiple beta regression analysis showed that gamification had an effect on self-reported process understanding: the subsample without gamification had a slightly higher self-reported process understanding than the subsample with (0.12; $p = .054$) (Tab. S934). However, here again, very little of the variance was explained ($R^2 = 0.55\%$).

3.4.3. Qualitative data for process understanding

The respondents receiving the nongamified treatment wrote more text that was coded as negative than positive about process understanding than those receiving the gamified treatment. Fewer respondents receiving the nongamified treatment without learning loop (31%) wrote comments coded as positive than respondents receiving the gamified treatment without learning loop (48%) ($\chi^2(1) = 12.914, p < .001$).

Among respondents writing comments, 8.3% understood that trade-

offs were unavoidable, 5.8% differentiated between alternatives and objectives, and 5.3% answered that they would reuse the method again. Reasons for not learning about the process were diverse: the method was too cognitively demanding (2.5%), they already knew the method (2%), applied it already (1%), used other methods (3%), or decided according to gut feeling (1.3%). The other methods mentioned were heuristics, AHP process, discussion with experts, SWOT analysis, “seven thinking steps,” cost-benefit analysis, simple utility analysis, risk analysis, and the FORDEC tool (S9.35).

More respondents receiving the gamified treatment (5.7%) made comments suggesting that they were thinking in terms of alternatives than those receiving the nongamified treatment (1.3%), ($\chi^2(1) = 12.201, p < .001$), for instance “options that one must weight against one another,” “Thinking in options,” “before a decision for the optimal objective, one should tradeoff between options”. More than half of the respondents receiving the nongamified treatment with learning loops (5.6%) reported that it was more cognitively demanding than all three other treatments (varying between 1 and 1.7), for instance “the survey completely exhausted me,” “Still difficult. One had to absorb and remember a relatively large amount of information in a concentrated way,” “The methods are interesting, but also a real challenge.”

3.5. Extraneous cognitive load

Overall, the respondents responded “rather no” to “some” extraneous cognitive load (3: some, 2: rather no on 5-point Likert scale; Tab.S941). Respondents with learning loops reported higher extraneous cognitive load than those without. The beta regression analysis showed that learning loop had an effect on the self-reported extraneous cognitive load: the subsample with learning loops reported statistically significantly higher extraneous cognitive load than the subsample without (0.25, $p = .0001$) (Tab.S942). Little variance was explained by the model ($R^2 = 2.26\%$). The qualitative data reported in Section 3.4 on process understanding corroborated these results.

3.6. Summary of the results

Some of our results met our expectations, although their practical relevance was limited in a number of cases (Table 9). For instance, the scores of the knowledge tests statistically significantly increased between the start and the end of the survey (confirming hypothesis 1a), but this increase was small (similarly to results in Aubert & Lienert, 2019). Moreover, the treatments explained very little of the overall variance in the dependent variables. Our results are in line with previous studies comparing designs of online surveys, including gamified ones, which have reported few differences in response patterns (e.g. Guin et al.,

2012). In our case, factual learning was higher with gamification and learning loops (confirming hypotheses H1b and H1c). Note, that our preliminary study with only 107 students had shown no statistically significant differences between the gamified survey and the control (Aubert & Lienert, 2019). However, the same respondents that received the learning loop treatment actually perceived their factual learning as lower (opposed to hypothesis H1c). Some preference construction occurred during the survey (confirming H2a), but significantly less for those receiving the nongamified treatment without learning loop than for the three other treatments (no hypothesis was formulated for the interactions between the treatments). Equally mixed results (obtained with different measure instruments than the ones used in this study) were reported for preference construction in Aubert and Lienert (2019). Process understanding regarding value-focused thinking, which had not been measured as such in earlier studies, occurred (confirming H3a) and was perceived as higher without gamification (no hypothesis was formulated). The extraneous cognitive load was perceived as significantly higher with learning loops, where respondents were asked to adjust their weights in cases of inconsistency (confirming H4). Overall, to some degree, all treatments enabled the occurrence of the three learning aspects that are prerequisites to decision-making, with slight differences between them. We discuss the results in the following section by reflecting on the improved operationalization of the measure instruments, the effects of the two features meant to enhance learning (gamification and learning loops), and how this study contributes to behavioral operational research (BOR).

4. Discussion

4.1. Measuring learning for decision-making

Before discussing the results, we reflect on our newly proposed comprehensive set of instruments to measure learning for decision-

making. These instruments are derived from a systematic operationalization and include novel measures for process understanding. Given our proposed conceptualization of learning for decision-making (which is discussed further in Section 4.4.2), we can recommend the further use of our measure instruments for factual learning, as well as for process understanding. Note that future work should adapt the knowledge test for factual learning to the topic of their decision, and control the test for validity previously to the experiment (e.g. Aubert et al., 2022).

Our measures for preference construction should be improved. Measuring preference construction remains a challenge for decisions analysts, because there is no right or wrong preference. Specifically, in our instrument, there were too few items in the performance test. It would be very important to develop new quantitative instruments to measure some kind of performance in preference construction. Also, measures for self-reported preference construction should be improved. Despite using three items found in the literature, the internal consistency between these was not high enough to aggregate them. In general, further work should focus on the internal consistency and external validation of the measure instruments. This requires careful experimental testing, as is standard in psychology.

4.2. Learning loops

Our previous experimental design did not allow us to investigate the effect of learning loops (Aubert & Lienert, 2019). In the present study, based on the literature, we had hypothesized that learning loops improve factual learning (H1c), preference construction (H2c), and process understanding (H3c), but increase the mental effort (H4). Results did not support the hypotheses on preference construction and process understanding (H2c and H3c). However, with learning loops, where respondents could adjust their weights in case of inconsistencies, respondents had a significantly higher wastewater knowledge score (supporting H1c), although the difference was small. With learning

Table 9

Summary of results. Hypotheses are summarized in Table 1. LL: learning loop. √: statistically significant confirming hypothesis. NA: no hypothesis was formulated (e.g. we did not know how learning loop and gamification would interact, neither did we foresee how gamification would affect process understanding). ×: hypothesis was not supported. (t): trend. n.s.: statistically insignificant difference between treatments.

Facets of learning	Data type	Measures	Results	Hypothesis supported or not
(1) Factual learning	Quantitative/ performance test	Delta score for knowledge test about wastewater objectives	Factual learning about objectives occurred Nongamified → lower score Without learning loop → lower score	√ H1a √ H1b √ H1c
		Delta score for knowledge test about wastewater alternatives	Nongamified → more Δ KS < 0 (final < initial KS) Factual learning about alternatives occurred Nongamified without learning loop → Δ KS < 0 (final < initial KS)	√ H1b √ H1a NA
	Quantitative/ self-reported	Self-reported confidence in the answers provided for the knowledge tests Self-reported factual learning	Confidence in answers increased Without learning loop → perceived higher Without learning loop → perceived higher	√ H1a × H1c × H1c
		Statements	In favor of gamified	√ (t) H1b
(2) Preference construction	Quantitative/ performance test	Test for preference construction (equivalent to consistency question) Patterns in the elicited weights	Nongamified without learning loop → lowest score n.s.	NA × H2b, H2c
		Self-reported preference construction	n.s.	× H2b, H2c
	Quantitative/ self-reported	Statements	Nongamified with learning loops → respondents voiced doubts about collected preferences Gamified → respondents had difficulty in formulating trade-offs and stated preferences for alternatives Gamified → respondents were concerned about population acceptance	NA (t) × (t) H2b NA (t)
			Process understanding occurred	√ H3a
(3) Process understanding	Quantitative/ performance test	Test for process understanding	Process understanding occurred	√ H3a
	Quantitative/self-reported	Self-reported process understanding	Nongamified → perceived higher	NA
	Qualitative	Statements	Nongamified treatment with learning loops → respondents found it more cognitively demanding Without learning loop → lower ECL	NA (t) √ H4
(4) Extraneous cognitive load (ECL)	Quantitative/self-reported	Self-reported ECL		√ H4

loops and nongamified survey, the performance scores for preference construction were also higher. Interestingly, respondents receiving the survey with learning loops perceived lower factual learning (contradicting H1c) and higher extraneous cognitive load (supporting H4), wrote more comments suggesting that the survey was complex or cognitively demanding (supporting H4), and, without gamification, also voiced doubts about the reliability of the expressed preferences. We thus observed a contradictory effect of learning loops: performance tests showed that factual learning somewhat increased, but the perception of this learning was lower.

The learning loops challenged many respondents, despite the improved interfaces, as indicated by the qualitative data and the higher extraneous cognitive load (ECL). Learning loops had been perceived as too complex in the qualitative assessments of our preliminary test (Aubert & Lienert, 2019) and we expected that it would be easier to deal with them with novel, improved instructions. Challenging tasks can positively influence learning (Hamari et al., 2016; Sheffler et al., 2022). However, to be positively perceived, the challenge should create a state of “flow” that matches the capabilities of the learners (Csikszentmihalyi, 1990). Otherwise, the challenge overwhelms, stresses, or bores learners (Csikszentmihalyi, 1990; Hamari et al., 2016). The higher ECL and lower perceived learning reported with learning loops suggest that our challenge was negatively perceived even though the factual learning test performance was often higher than without learning loop. Future work could investigate the disparity between measured and perceived learning also reported in education literature (e.g. Deslauriers, McCarty Logan, Miller, Callaghan & Kestin, 2019), and the associated causalities, for instance using the theory of flow (Csikszentmihalyi, 1990).

We maintain that learning loops with consistency checks are important and good practice (Anderson & Clemen, 2013; Payne et al., 2006). They have been recommended on the basis of practical applications (e.g. Hobbs & Meier, 1994; Martin, 2021; Marttunen & Hämäläinen, 2008), and more recently of a computational experiment (Lahtinen, Hämäläinen & Jenytin, 2020). Our learning loops can be improved in two ways. The first is to better assist respondents. Personalized feedback and instructions could guide users lost in learning loops, for example with a chatbot using artificial intelligence (AI). AI has been used to develop learner-centered online learning platforms (Merriam & Baumgartner, 2020) and to elicit stakeholders’ preferences (Toffano et al., 2022). The second way is to simplify the learning loops by designing less cognitively demanding consistency checks. Recently demonstrated for discrete choice modeling, simple models increased problem understanding (Tako, Tsiptsias & Robinson, 2020). For instance, learning loops could ask the respondents to select correct statements about their preferences instead of requiring a numerical answer. For example, “In my opinion, improving Objective X from worst to best is more important than improving Objective Y from worst to best.” Alternatively, objectives could be directly ranked. In sum, the learning loops are good, but should be simplified. This follows a recent proposition of cognitive psychology scientists, who also argue that there was no added-value of complex models that may be more accurate compared to more simple and understandable models that can easily be used in practice (e.g. “fast and frugal heuristics tree”) (Katsikopoulos, Ozgur, Buckmann & Gigerenzer, 2020). Moreover, although challenges can improve learning (Hamari et al., 2016; Sheffler et al., 2022), one BOR study found that it was ineffective to debias overprecision in estimates (Ferretti et al., 2022). Future work should explore these think-harder strategies for preference construction further.

Finally, other ways might be found to enhance (and assess) preference construction. A recent online example, also using cognitive dissonance, proposed to activate values (Philpot et al., 2022). These authors prompted respondents to compare their preferences for alternatives, and their values. Our results support future attempts using learning loops if they keep the ECL low. Research could assess their effectiveness and underlying mechanisms.

4.3. Gamification

We had hypothesized that gamification improves factual learning (H1b) and preference construction (H2b). Results supported the hypotheses on factual learning (H1b) but not on preference construction (H2b). In our previous studies with smaller student samples, we did not observe statistically significant differences in factual learning between the gamified and control treatments (Aubert & Lienert, 2019). Additionally, in the earlier studies, we did not observe differences in preference construction (note that the measure instruments differed as we used improved, novel instruments in the present study). Overall, both gamified and nongamified treatments were neutral to good in all three learning aspects, although the treatments explained little of the variance. As expected, gamification led to somewhat higher knowledge scores for understanding wastewater objectives (supporting hypothesis H1b), and, rather unexpectedly, to a lower perceived understanding of value-focused thinking (no hypothesis was formulated). Respondents with gamification wrote more comments. These indicated that respondents learned facts about wastewater management (supporting H1b) and realized that making trade-offs is necessary but difficult (contradicting H2b). Moreover, respondents voiced preferences for alternatives and concerns about the population’s acceptance of decentralized wastewater alternatives. Finally, the dropout rate was higher for the gamified treatment than the nongamified one. Technical issues caused some dropouts, but some respondents also emailed us that they stopped the survey because its gamified format trivialized the important topic. These innovative game elements included the storyline with the challenge of winning the next election, an avatar chosen by the respondent who meets gender-neutral nonplayer characters, and ambient music among others (Section 2.2 and S4.3). The perceived trivialization of the wastewater topic criticized by some respondents may be due to using animal characters (see right panel in Fig. 3) rather than people as in the earlier prototypes (Aubert & Lienert, 2019; Aubert et al., 2022). After discussion with game designers, we had decided for animals as they are gender-neutral. Appropriate types of characters, avatar and nonplayer, for serious games could be a future research topic (e.g. Kim, Lee & Chung, 2023).

The high dropout rate and the higher number of voluntary qualitative comments are in line with the existing literature on gamified surveys (e.g. Bailey et al., 2015; Keusch & Zhang, 2015). The dropout rate in the gamified treatment might have biased our study: only those open to the new format could have answered the feedback questions. However, our samples with and without gamification are representative of the targeted population in gender and age (S5–7). Thus, we assume that the preferences are representative of our targeted population. Future studies could let the respondents choose between a gamified and a nongamified survey and investigate the respondents’ profiles for characteristics that may determine their choice. These may include the attitude to games (Guin et al., 2012) and individual motivational orientations. According to causality orientation theory (Loughrey & Broin, 2018; Ryan & Deci, 2017), individuals react differently to motivational stimuli. For instance, an autonomy-oriented person may be frustrated by external stimuli, whereas the same stimuli may motivate a controlled-oriented person. Moreover, personality traits could influence whether people like gamification, such as the Big Five (openness, conscientiousness, extraversion, agreeableness, and neuroticism) (Triantoro, Gopal, Benbunan-Fich & Lang, 2019, 2020).

For those open to this new format, gamification successfully created awareness about the complexity of wastewater management (supporting hypothesis H1b) and evoked a positive attitude. However, this positive attitude might have biased the feedback, because those who disliked the format were underrepresented (i.e. they dropped out). We observed factual learning (supporting hypothesis H1b), but it was limited. Additionally, the qualitative feedback for factual learning and preference construction clearly indicated that respondents realized that decisions about wastewater management required making trade-offs, but they

found it difficult to make these trade-offs. Our mixed results thus allow us to state that gamification led to successful awareness raising rather than successful learning.

Another unexpected but, we think, positive result is that respondents receiving the gamified treatment voiced concerns about the population's acceptance of decentralized wastewater alternatives. They highlighted that such decisions require broad information, should satisfy divergent interests, and that the best compromise should be found. We interpreted this as social learning, or learning about diverse worldviews (den Haan & van der Voort, 2018; Reed et al., 2010). This could have been primed by the narrative of the gamified survey. It included meeting citizens who advocated a single objective, and respondents playing candidates to represent these citizens at the municipal council. Some respondents forgot that the aim of the survey was to collect their own preferences to inform a decision. Rather, they interpreted their role as that of a mediator. This is a known effect: the game activity can distract from the main task (Keusch & Zhang, 2015). This may also have caused the lower perceived process understanding with gamification. Follow-up studies could focus on this transfer of learning from the "magic circle" of the game (Huizinga, 1949) to the real-world task.

Some counterarguments about gamification also warrant consideration. First, our gamification was strongly constrained by the norms of the swing and trade-off weight elicitation methods (Eisenführ et al., 2010; von Winterfeldt & Edwards, 1986). Considering simpler methods to elicit weights would be interesting, and the game designers were also keen to simplify weight elicitation (S4.4). Follow-up attempts to gamify weight elicitation for MCDA could ask respondents simply to rank objectives. We would then calculate weights from rankings (e.g. Riabacke et al., 2012; Roberts & Goodwin, 2002). Second, developing our gamified survey required substantial time and money and relied on game designers. Although our gamified survey concept is easily adaptable, new texts and artwork will be needed for the next application case, which will consume resources. Before gamifying a survey, one should weigh the pros and cons, which can be context specific.

Finally, the novel nongamified survey developed here presents some practical advantages. First, the admin interface enables new weight elicitation surveys to be created quickly for other decisions and the weight elicitation process meets the standards. Second, the nongamified survey can also support interviews and facilitated workshops, enabling display of the weights. Third, the code is open source and available: the software for the nongamified survey only needs to be deployed on a server (Aubert & Masson, 2021).

4.4. Insights for OR

4.4.1. OR and e-participation

Our four survey treatments successfully raised awareness in a sample representative of the Swiss population about a topical decision. Respondents learnt about an unfamiliar topic, constructed and shared preferences, and understood that deciding about wastewater management requires making difficult trade-offs between many objectives. The interfaces informed respondents effectively. Moreover, the survey can support co-deciding, if it is used before and/or after more deliberative types of participation, such as stakeholder workshops and citizens' fora. In sum, our online survey could support the three main aims of public participation in decision-making: informing, consulting, and co-deciding (Arnstein, 1969; de Gooyert, Rouwette, van Kranenburg & Freeman, 2017). The decision-makers should clarify the aim of participation at the start and inform respondents about the decision problem and methods used (Linkov et al., 2006).

Our online survey supported a detailed, in-depth understanding of the decision problem and enabled many citizens to participate. This aligns with the paradigms of deep and broad participation (Gregory, Satterfield & Hasell, 2016; Papadopoulos & Warin, 2007). However, the digital divide remains a reality, with disparities in internet access and use (Merriam & Baumgartner, 2020). Thus, surveys such as ours

complement rather than replace existing participatory decision-making processes (French & Argyris, 2018; Gregory et al., 2016; Philpot et al., 2022).

Our nongamified survey can be used in other settings. It can support an interactive group weight elicitation workshop, for instance with visualization, or individual interviews. In these examples, a decision analyst would facilitate the process. To date, the novel interface does not support asynchronous, autonomous use by a group. Future development could add features to display weights of multiple stakeholders and to automatically support a fair group process, for instance following the group Delphi method (Niederberger & Renn, 2018).

Originally, we developed the survey to enable individuals to answer without guidance from a decision analyst. Our results displayed high interpersonal variability, with fewer significant differences between the treatments than expected. It is intuitively apparent that different types of games appeal to different types of people (Guin et al., 2012; Triantoro et al., 2019). Additionally, we observed that the learning loops also triggered people differently, possibly depending on their cognitive capabilities. A recent meta-analysis, based on 30 articles on risk preferences, confirmed that cognitive abilities can lead to erroneous choice behaviors (Mechera-Ostrovsky, Heinke, Andraszewicz & Rieskamp, 2022). Consequently, high interpersonal variability challenges the design of an interface meant for a broad public. For this reason, we suggest that citizens could choose their preferred interface, and further studies should examine individual characteristics. Further studies should also test different learning loops and continue to identify the benefits and drawbacks of surveytainment (Kostyk et al., 2019).

4.4.2. Learning and decision-making

Decision analysis, including modeling and structured decision-making, has long been described as an iterative learning processes (e.g. Belton & Elder, 1994). McDaniels and Gregory (2004) suggest capturing learning as a separate fundamental objective in structured decision-making. In the BOR literature, learning as a dependent variable is mostly studied at the individual level (Monks et al., 2014, 2016; Tako et al., 2020; Thompson et al., 2016) despite the call to follow the rich literature on group model building of system dynamics (Franco et al., 2021). Our study also focused on learning for individuals but targeted laypeople. Such affected stakeholders can be represented in a group decision-making process, but it is rare except in community OR (e.g. Midgley et al., 2018).

Based on learning (e.g. Mezirow, 2000) and decision analysis literature (e.g. Belton & Elder, 1994), we conceptualized learning for decision-making as comprising factual learning, preference construction, and process understanding. Our novel assessment method proved satisfactory, except for preference construction. Unexpectedly, the qualitative feedback in the gamified version of the present study highlighted that social learning is a fourth facet of learning that we had not considered. Further studies should refine our assessment method to measure preference construction and to include social learning. Our results confirmed the complementarity of performance tests, self-reported assessments, and qualitative data (den Haan & van der Voort, 2018; Franco et al., 2021; Hamari et al., 2016). Complementary data revealed that perceived learning can differ from learning measured by performance tests, as observed here and in previous studies (Aubert & Lienert, 2019; Deslauriers et al., 2019). It also revealed interesting questions. As example: how can we explain that the difference between the final and initial knowledge scores was more often negative for alternatives than for objectives? Can this be only explained by tiredness (the questions about alternatives came after those about objectives)? Or did we emphasize objectives so strongly (following value-focused thinking) that alternatives appeared less important? We propose testing our assessment method in other contexts, for instance in decision-making workshops. This would provide a baseline for comparing our interfaces. Finally, to better unravel critical learning incidents, other methods are more suitable, such as think-aloud

protocols, video analysis, click-data, and eye-tracking (Franco et al., 2021; Hamari et al., 2016). We encourage future BOR studies to continue investigating learning as a process.

Understanding the learning process will enable improving the design of OR interventions. Learning needs time, as respondents need to process the information (Gregory et al., 2012). Several respondents reported in the qualitative feedback that they would need to be further informed or would look for complementary information sources. We propose that future development could decouple the informing part from preference elicitation. The informing part could include links and references to information sources. Additionally, the preference elicitation part should ensure that the facts are always easily accessible. It could also include small training and practice tasks for swing and trade-off weight elicitation to enhance process understanding (Anderson & Clemen, 2013). Moreover, the respondents should be given the opportunity to take a break from the survey (which was not possible in our case), and could be offered a chance to review and revise their preferences after a set time. Our novel structured assessment method presented in this paper enables comparing all these development options and identifying which most improve the respondents' learning.

As drawback, these developments are resource intensive. Although learning requires time, the survey length is critical. Long surveys increase dropout rates. Good practice recommends a maximum of 20 questions, per stage if multistage, and maximally 13 min to complete the survey (Bailey et al., 2015). However, if respondents think that the data are important, they seem to accept a higher survey burden (Guin et al., 2012). Our survey was long and complicated, and should be shortened. Simpler consistency checks could be tested for the learning loops. The relevance of consistency checks to enhance preference construction should be clarified, as the think-harder strategy seemed to raise doubt and distrust among some respondents. Finally, the interfaces should be adapted for a range of devices, because people increasingly use smaller screens, including cell phones. Because of the costs of such developments, we highly recommend OR researchers and practitioners to first investigate how to best enhance learning.

5. Conclusion and outlook

Individual learning for decision-making is topical for BOR. In our study, we focused on laypeople, investigating whether learning for decision-making occurred when citizens answered specifically designed online weight elicitation surveys. In addition, we investigated the effect of two features meant to increase learning – gamification and learning loops. We used a comprehensive, innovative set of instruments to measure learning for decision-making, derived from a systematic operationalization.

Our conceptualization of learning for decision-making comprised factual learning, preference construction and process understanding. Because we targeted laypeople, we interpreted process understanding as understanding value-focused thinking, i.e. differentiating objectives from alternatives and understanding that the relative importance of objectives is used to derive a ranking of alternatives. Although being more comprehensive than previous attempts, our framework missed social learning, which was unexpectedly triggered by the storyline of the gamified interface. Future OR work investigating learning for decision-making should consider all four learning facets.

Using newly-designed interfaces for eliciting weights online, we successfully raised awareness about a public decision concerning wastewater management. Most respondents learnt something about the facts, their own preferences, and value-focused thinking. Gamification and learning loops enhanced factual learning, however learning loops decreased the perception of process understanding. Our results also revealed high interpersonal variability, which is really challenging if we aim at designing user-centered interfaces. In future research, we could study effects of letting the respondents choose the survey format (e.g. gamified

or not).

The gamification feature was useful: it enhanced topical awareness and social learning. However, drop-out was high, the practical relevance of the differences to the nongamified interface was low, and developing the gamified version was resource-intensive. For online weight elicitation, based on the present study, we recommend adding a few game elements to the nongamified interface, rather than developing a fully-fledged gamified interface.

The learning loop feature was useful to increase factual learning, but was clearly cognitively demanding and lowered the perceived learning on all three facets. We propose using “think-harder” strategies to further investigate this difference between tested and perceived learning. OR researchers interested in online weight elicitation have to develop tools that assist the respondents more effectively and may consider simplified procedures. We could investigate whether simply ranking the objectives to derive weights is sufficient. We could compare the effects of our learning loops with simpler loops, since we found that thinking harder does not necessarily improve learning. Most likely, we should consider the aim of participation: raising awareness does not require the same tools as co-deciding.

Finally, we newly operationalized the measure instruments of learning for decision-making with complementary types of data (quantitative performance test and self-assessment, qualitative self-assessment). We used this new assessment framework to compare our four interfaces. Further development should revise the measures for preference construction, and add performance tests and self-reported measures for social learning, thus transcending individual learning. We insist on using complementary data types: qualitative feedback provided insights that helped explain the quantitative results. Other instruments such as think-aloud protocols may well be appropriate if the aim is to unravel the learning processes, not only the learning outcomes. OR researchers could use our learning for decision-making framework to measure learning in traditional settings, such as interviews and workshops. This baseline would allow comparison with innovative tools such as the ones presented here.

To conclude, despite many new open questions, our results with these innovative interfaces proved promising: they successfully raised awareness among many affected citizens in public decision-making.

Funding

Ambizione grant #173973 to Aubert from the Swiss National Science Foundation

CRediT authorship contribution statement

Alice H. Aubert: Conceptualization, Methodology, Writing – original draft, Visualization, Supervision, Project administration, Funding acquisition. **Sara Schmid:** Formal analysis, Investigation, Data curation. **Judit Lienert:** Methodology, Writing – review & editing.

Acknowledgements

We thank all respondents, in particular those who voluntarily provided additional detailed qualitative feedback. We warmly thank Philipp Beutler, Harald von Waldow, and Stuart Dennis from Eawag, Antoine Masson from Youmi, the teams from Entrée de Jeux and OpinionGames (Software), and the team from Intervista (market research company). Finally, we thank the three reviewers for their constructive feedback.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at [doi:10.1016/j.ejor.2023.10.031](https://doi.org/10.1016/j.ejor.2023.10.031).

References

- Adcock, A. (2012). Cognitive dissonance in the learning process. In N. M. Seel (Ed.), *Encyclopedia of the sciences of learning* (pp. 2182–2184). Springer. <https://doi.org/10.1007/978-1-4419-1428-6>.
- Anderson, R. M., & Clemen, R. (2013). Toward an improved methodology to construct and reconcile decision analytic preference judgments. *Decision Analysis*, 10(2), 121–134. <https://doi.org/10.1287/deca.2013.0268>
- Argyris, C. (1978). *Organizational learning: A theory of action perspective*. Reading: Addison-Wesley.
- Arnstein, S. R. (1969). A ladder of citizen participation. *Journal of the American Institute of Planners*, 35(4), 216–224. <https://doi.org/10.1080/01944366908977225>
- Arvai, J. L., Gregory, R., & McDaniels, T. L. (2001). Testing a structured decision approach: Value-focused thinking for deliberative risk communication. *Risk Analysis*, 21(6), 1065–1076. <https://doi.org/10.1111/0272-4332.216175>
- 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. <https://doi.org/10.1016/j.envsoft.2018.03.023>
- Aubert, A. H., Esculier, F., & Lienert, J. (2020). Recommendations for online elicitation of swing weights from citizens in environmental decision-making. *Operations Research Perspectives*, 7(100156). <https://doi.org/10.1016/j.orp.2020.100156>
- Aubert, A. H., & Lienert, J. (2019). Gamified online survey to elicit citizens' preferences and enhance learning for environmental decisions. *Environmental Modelling & Software*, 111, 1–12. <https://doi.org/10.1016/j.envsoft.2018.09.013>
- Aubert, A. H., Lienert, J., & von Helversen, B. (2022). Gamified environmental multi-criteria decision analysis: Information on objectives and range insensitivity bias. *International Transactions in Operational Research*, 1–33. <https://doi.org/10.1111/itor.13206>
- Aubert, A. H., & Masson, A. (2021). *Software to elicit weights for multi-criteria decision analysis*. Eawag: Swiss Federal Institute of Aquatic Science and Technology. <https://doi.org/10.25678/00056J>. In (Version 1.0).
- Aubert, A. H., Schmid, S., Beutler, P., & Lienert, J. (2022). Innovative online survey about sustainable wastewater management: What young Swiss citizens know and value. *Environmental Science & Policy*, 137, 323–335. <https://doi.org/10.1016/j.envsci.2022.08.018>
- 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. <https://doi.org/10.2501/IJMR-2015-003>
- Belton, V., & Elder, M. D. (1994). Decision support systems: Learning from visual interactive modelling. *Decision Support Systems*, 12(4), 355–364. [https://doi.org/10.1016/0167-9236\(94\)90052-3](https://doi.org/10.1016/0167-9236(94)90052-3)
- Bessette, D. L., Campbell-Arvai, V., & Arvai, J. (2016). Expanding the reach of participatory risk management: Testing an online decision-aiding framework for informing internally consistent choices. *Risk Analysis*, 36(5), 992–1005. <https://doi.org/10.1111/risa.12481>
- Beutler, P., & Lienert, J. (2019). Project "Decision-making for the transition from central to novel wastewater infrastructure systems" <https://www.eawag.ch/en/departments/ess/projects/decision-making-for-wastewater-infrastructure-transitions/> (retrieved on 11.10.2019).
- Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education*, 32(3), 347–364. <http://www.jstor.org/stable/3448076>.
- Clemen, R. T. (1991). Making hard decisions: An introduction to decision analysis. PWS-Kent Publ. Co.
- Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience*. Harper & Row.
- Danilenko, E. P. (2010). *PhD dissertation* (p. 120). The faculty of the graduate school of the university of Minnesota.
- de Gooyert, V., Rouwette, E., van Kranenburg, H., & Freeman, E. (2017). Reviewing the role of stakeholders in operational research: A stakeholder theory perspective. *European Journal of Operational Research*, 262(2), 402–410. <https://doi.org/10.1016/j.ejor.2017.03.079>
- den Haan, R. J., & van der Voort, M. (2018). On evaluating social learning outcomes of serious games to collaboratively address sustainability problems: A literature review. *Sustainability*, 10(12), 4529. <http://www.mdpi.com/2071-1050/10/12/4529>.
- Deslauriers, L., McCarty Logan, S., Miller, K., Callaghan, K., & Kestin, G. (2019). Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. *Proceedings of the National Academy of Sciences*, 116(39), 19251–19257. <https://doi.org/10.1073/pnas.1821936116>
- Deterding, S., Dixon, D., Khaled, R., & Nacke, L. E. (2011). *From game design elements to gamefulness: Defining "Gamification"*. Tampere, Finland: MindTrek'11.
- Dichev, C., & Dicheva, D. (2017). Gamifying education: What is known, what is believed and what remains uncertain: A critical review [Review]. *International Journal of Educational Technology in Higher Education*, 14(36), Article 9. <https://doi.org/10.1186/s41239-017-0042-5>
- Eisenführ, F., Weber, M., & Langer, T. (2010). *Rational decision making*. Germany: Springer Berlin Heidelberg. <http://www.rationalesentscheiden.de/>.
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41(4), 1149–1160. <https://doi.org/10.3758/BRM.41.4.1149>
- Ferretti, V., Montibeller, G., & von Winterfeldt, D. (2022). Testing the effectiveness of debiasing techniques to reduce overprecision in the elicitation of subjective continuous probability distributions. *European Journal of Operational Research*. <https://doi.org/10.1016/j.ejor.2022.04.008>
- Ferretti, V., Pluchinotta, I., & Tsoukiàs, A. (2019). Studying the generation of alternatives in public policy making processes. *European Journal of Operational Research*, 273(1), 353–363. <https://doi.org/10.1016/j.ejor.2018.07.054>
- Franco, L. A., Hämäläinen, R. P., Rouwette, E. A. J. A., & Leppänen, I. (2021). Taking stock of behavioural OR: A review of behavioural studies with an intervention focus. *European Journal of Operational Research*, 293(2), 401–418. <https://doi.org/10.1016/j.ejor.2020.11.031>
- Franco, L. A., & Lord, E. (2011). Understanding multi-methodology: Evaluating the perceived impact of mixing methods for group budgetary decisions. *Omega*, 39(3), 362–372. <https://doi.org/10.1016/j.omega.2010.06.008>
- French, S., & Argyris, N. (2018). Decision analysis and political processes. *Decision Analysis*, 15(4), 208–222. <https://doi.org/10.1287/deca.2018.0374>
- 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. <http://www.structureddecisionmaking.org/>.
- 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. <https://doi.org/10.1073/pnas.1508896113>
- Guin, T. D.-L., Baker, R., Mechling, J., & Ruyle, E. (2012). Myths and realities of respondent engagement in online surveys. *International Journal of Market Research*, 54(5), 613–633. <https://doi.org/10.2501/IJMR-54-5-613-633>
- Haag, F., Zürcher, S., & Lienert, J. (2019). Enhancing the elicitation of diverse decision objectives for public planning. *European Journal of Operational Research*. <https://doi.org/10.1016/j.ejor.2019.06.002>
- Hamari, J., Shernoff, D. J., Rowe, E., Coller, B., Asbell-Clarke, J., & Edwards, T. (2016). Challenging games help students learn: An empirical study on engagement, flow and immersion in game-based learning. *Computers in Human Behavior*, 54, 170–179. <https://doi.org/10.1016/j.chb.2015.07.045>
- Harms, J., Biegler, S., Wimmer, C., Kappel, K., & Grechenig, T. (2015). Gamification of online surveys: Design process, case study, and evaluation. In J. Abascal, S. Barbosa, M. Fetter, T. Gross, P. Palanque, & M. Winckler (Eds.), *Human-Computer interaction – interact 2015* (pp. 219–236). Springer International Publishing.
- Hobbs, B. F., & Meier, P. M. (1994). Multicriteria methods for resource planning: An experimental comparison. *IEEE Transactions on Power Systems*, 9(4), 1811–1817. <https://doi.org/10.1109/59.331435>
- Hoffmann, S., Feldmann, U., Bach, P. M., Binz, C., Farrelly, M., Frantzeskaki, N., Hiessl, H., Inauen, J., Larsen, T. A., Lienert, J., Londong, J., Lüthi, C., Maurer, M., Mitchell, C., Morgenroth, E., Nelson, K. L., Scholten, L., Truffer, B., & Udert, K. M. (2020). A research agenda for the future of urban water management: Exploring the potential of nongrid, small-grid, and hybrid solutions. *Environmental Science & Technology*, 54(9), 5312–5322. <https://doi.org/10.1021/acs.est.9b05222>
- Huizinga, J. (1949). *Homo Ludens: A study of the play-element in culture*. Routledge & Kegan Paul Ltd.
- Johnson, M. P., Midgley, G., & Chichirau, G. (2018). Emerging trends and new frontiers in community operational research. *European Journal of Operational Research*, 268(3), 1178–1191. <https://doi.org/10.1016/j.ejor.2017.11.032>
- Kaplan, E. H. (2008). Adventures in policy modeling! Operations research in the community and beyond. *Omega*, 36(1), 1–9. <https://doi.org/10.1016/j.omega.2005.07.012>
- Kasurinen, J., & Knutas, A. (2018). Publication trends in gamification: A systematic mapping study. *Computer Science Review*, 27, 33–44. <https://doi.org/10.1016/j.cosrev.2017.10.003>
- Katsikopoulos, K. V. S., Ozgur, Buckmann, Marcus, & Gigerenzer, Gerd (2020). *Classification in the wild: The science and art of transparent decision making*. The MIT Press.
- Keeney, R. L. (1992). *Value-Focused thinking: A path to creative decision-making*. Cambridge, Massachusetts: Harvard University Press.
- Keeney, R. L., & Raiffa, H. (1976). *Decisions with multiple objectives: Preferences and value tradeoffs*. Wiley.
- Keusch, F., & Zhang, C. (2015). A review of issues in gamified surveys. *Social Science Computer Review*, 35(2), 147–166. <https://doi.org/10.1177/0894439315608451>
- Kim, D. Y., Lee, H. K., & Chung, K. (2023). Avatar-mediated experience in the metaverse: The impact of avatar realism on user-avatar relationship. *Journal of Retailing and Consumer Services*, 73, Article 103382. <https://doi.org/10.1016/j.jretconser.2023.103382>
- Koivisto, J., & Hamari, J. (2019). The rise of motivational information systems: A review of gamification research. *International Journal of Information Management*, 45, 191–210. <https://doi.org/10.1016/j.ijinfomgt.2018.10.013>
- Kostyk, A., Zhou, W., & Hyman, M. R. (2019). Using surveytainment to counter declining survey data quality. *Journal of Business Research*, 95, 211–219. <https://doi.org/10.1016/j.jbusres.2018.10.024>
- Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys [https://doi.org/10.1002/acp.2350050305]. *Applied Cognitive Psychology*, 5(3), 213–236. <https://doi.org/10.1002/acp.2350050305>
- Lahtinen, T. J., Hämäläinen, R. P., & Jenytn, C. (2020). On preference elicitation processes which mitigate the accumulation of biases in multi-criteria decision analysis. *European Journal of Operational Research*, 282(1), 201–210. <https://doi.org/10.1016/j.ejor.2019.09.004>
- Landers, R. N., Auer, E. M., Collmus, A. B., & Armstrong, M. B. (2018). Gamification science, its history and future: Definitions and a research agenda. *Simulation & Gaming*, 49(3), 315–337. <https://doi.org/10.1177/1046878118774385>
- Larsen, T. A., Hoffmann, S., Lüthi, C., Truffer, B., & Maurer, M. (2016). Emerging solutions to the water challenges of an urbanizing world. *Science (New York, N.Y.)*, 352(6288), 928–933. <https://doi.org/10.1126/science.aad8641>
- Leppink, J., Paas, F., Van der Vleuten, C. P. M., Van Gog, T., & Van Merriënboer, J. J. G. (2013). Development of an instrument for measuring different types of cognitive

- load. *Behavior Research Methods*, 45(4), 1058–1072. <https://doi.org/10.3758/s13428-013-0334-1>
- Leppink, J., Paas, F., van Gog, T., van der Vleuten, C. P. M., & van Merriënboer, J. J. G. (2014). Effects of pairs of problems and examples on task performance and different types of cognitive load. *Learning and Instruction*, 30, 32–42. <https://doi.org/10.1016/j.learninstruc.2013.12.001>
- Lichtenstein, S., & Slovic, P. (2006). *The construction of preference*. New York: Cambridge University Press.
- 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. <https://doi.org/10.1016/j.ejor.2016.03.010>
- Linkov, I., Satterstrom, F. K., Kiker, G., Batchelor, C., Bridges, T., & Ferguson, E. (2006). From comparative risk assessment to multi-criteria decision analysis and adaptive management: Recent developments and applications. *Environment International*, 32(8), 1072–1093. <https://doi.org/10.1016/j.envint.2006.06.013>
- Loughrey, K., & Broin, D.O. (.2018., 15-17 Aug. 2018). Are We Having Fun Yet? Misapplying Motivation to Gamification. 2018 IEEE Games, Entertainment, Media Conference (GEM).
- Martin, D. M. (2021). Navigating inconsistent preferences: A multimethod approach to making informed decisions. *Conservation Science and Practice*, 3(8), e469. <https://doi.org/10.1111/csp2.469>
- Marttunen, M., & Hämäläinen, R. P. (2008). The decision analysis interview approach in the collaborative management of a large regulated water course [journal article]. *Environmental Management*, 42(6), 1026–1042. <https://doi.org/10.1007/s00267-008-9200-9>
- McDaniels, T. L., & Gregory, R. (2004). Learning as an objective within a structured risk management decision process. *Environmental Science & Technology*, 38(7), 1921–1926. <https://doi.org/10.1021/es0264246>
- Mechera-Ostrovsky, T., Heinke, S., Andraszewicz, S., & Rieskamp, J. (2022). Cognitive abilities affect decision errors but not risk preferences: A meta-analysis. *Psychonomic Bulletin & Review*. <https://doi.org/10.3758/s13423-021-02053-1>
- Merriam, S. B., & Baumgartner, L. M. (2020). *Learning in adulthood: A comprehensive guide* (4th ed.). Jossey-Bass.
- Mezirow, J. (2000). *Learning as transformation: Critical perspectives on a theory in progress*. San Francisco: Jossey Bass.
- Midgley, G., Cavana, R. Y., Brocklesby, J., Foote, J. L., Wood, D. R. R., & Ahuriri-Driscoll, A. (2013). Towards a new framework for evaluating systemic problem structuring methods. *European Journal of Operational Research*, 229(1), 143–154. <https://doi.org/10.1016/j.ejor.2013.01.047>
- Midgley, G., Johnson, M. P., & Chichirau, G. (2018). What is community operational research? *European Journal of Operational Research*, 268(3), 771–783. <https://doi.org/10.1016/j.ejor.2017.08.014>
- Monks, T., Robinson, S., & Kotiadis, K. (2014). Learning from discrete-event simulation: Exploring the high involvement hypothesis. *European Journal of Operational Research*, 235(1), 195–205. <https://doi.org/10.1016/j.ejor.2013.10.003>
- 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. <https://doi.org/10.1016/j.ejor.2015.08.037>
- Montbeller, G., & von Winterfeldt, D. (2015). Cognitive and motivational biases in decision and risk analysis. *Risk Analysis*, 35(7), 1230–1251. <https://doi.org/10.1111/risa.12360>
- Niederberger, M., & Renn, O. (2018). *Das gruppendelphi-verfahren: Vom konzept bis zur anwendung* (1st ed.). Springer Fachmedien Wiesbaden. <https://doi.org/10.1007/978-3-658-18755-2>
- Papadopoulos, Y., & Warin, P. (2007). Are innovative, participatory and deliberative procedures in policy making democratic and effective? *European Journal of Political Research*, 46(4), 445–472. <https://doi.org/10.1111/j.1475-6765.2007.00696.x>
- Payne, J. W., Bettman, J. R., & Schkade, D. A. (2006). Measuring constructed preferences: Towards a building code. In S. Lichtenstein, & P. Slovic (Eds.), *The construction of preference* (pp. 629–652). New York: Cambridge University Press.
- Philpot, S. L., Philpot, C. T., Hipel, K. W., & Johnson, P. A. (2022). Proposal, design, and evaluation of a values-based online survey. *EURO Journal on Decision Processes*, Article 100015. <https://doi.org/10.1016/j.ejdp.2022.100015>
- Plass, J. L., Homer, B. D., & Kinzer, C. K. (2015). Foundations of game-based learning. *Educational Psychologist*, 50(4), 258–283. <https://doi.org/10.1080/00461520.2015.1122533>
- R Development Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. URL <http://www.R-project.org/>.
- Reed, M. S., Evely, A. C., Cundill, G., Fazey, I., Glass, J., Laing, A., Newig, J., Parrish, B., Prell, C., Raymond, C., & Stringer, L. C. (2010). What is Social Learning? *Ecology and Society*, 15(4), Article r1. <https://doi.org/10.5751/ES-03564-1504r01>
- Riabacke, M., Danielson, M., & Ekenberg, L. (2012). State-of-the-art prescriptive criteria weight elicitation. *Advances in Decision Sciences*, 2012, 1-24, Article 276584. <https://doi.org/10.1155/2012/276584>
- Ríos Insua, D.R., & French, S. (2010). e-Democracy: A group decision and negotiation perspective (D. & R. Insua & S. French, Eds.). Springer. <https://doi.org/10.1007/978-90-481-9045-4>.
- Roberts, R., & Goodwin, P. (2002). Weight approximations in multi-attribute decision models. *Journal of Multi-Criteria Decision Analysis*, 11(6), 291–303. <https://doi.org/10.1002/mcda.320>
- 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. <https://doi.org/10.1016/j.chb.2017.03.048>
- Ryan, R., & Rigby, C. S. (2019). Motivational foundations of game-based learning (chap.6). In J. L. Plass, R. & E. Mayer, & B. D. Homer (Eds.), *Handbook of game-based learning* (pp. 153–176).
- Ryan, R. M., & Deci, E. L. (2017). *Self-determination theory: Basic psychological needs in motivation, development, and wellness*. New York: The Guilford Press.
- Sheffler, P., Rodriguez, T. M., Cheung, C. S., & Wu, R. (2022). Cognitive and metacognitive, motivational, and resource considerations for learning new skills across the lifespan [https://doi.org/10.1002/wcs.1585]. *WIREs Cognitive Science*, 13(2), e1585. <https://doi.org/10.1002/wcs.1585>
- Subhash, S., & Cudney, E. A. (2018). Gamified learning in higher education: A systematic review of the literature. *Computers in Human Behavior*, 87, 192–206. <https://doi.org/10.1016/j.chb.2018.05.028>
- Tako, A. A., Tsiptsias, N., & Robinson, S. (2020). Can we learn from simplified simulation models? An experimental study on user learning. *Journal of Simulation*, 14(2), 130–144. <https://doi.org/10.1080/17477778.2019.1704636>
- 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. <https://doi.org/10.1016/j.ejor.2015.09.048>
- Toffano, F., Garraffa, M., Lin, Y., Prestwich, S., Simonis, H., & Wilson, N. (2022). A multi-objective supplier selection framework based on user-preferences. *Annals of Operations Research*, 308(1), 609–640. <https://doi.org/10.1007/s10479-021-04251-5>
- Triantoro, T., Gopal, R., Benbunan-Fich, R., & Lang, G. (2019). A comparison of a gamified survey with a traditional online survey method. *International Journal of Information Management*, 49, 242–252. <https://doi.org/10.1016/j.ijinfomgt.2019.06.001>
- Triantoro, T., Gopal, R., Benbunan-Fich, R., & Lang, G. (2020). Personality and games: Enhancing online surveys through gamification. *Information Technology and Management*, 21(3), 169–178. <https://doi.org/10.1007/s10799-020-00314-4>
- van Roy, R., & Zaman, B. (2017). Why gamification fails in education and how to make it successful: Introducing nine gamification heuristics based on self-determination theory. In O. A. E. Ma M. (Ed.), *Serious games and edutainment applications*. Cham: Springer. https://doi.org/10.1007/978-3-319-51645-5_22.
- Vieira, A. C. L., Oliveira, M. D., & Bana e Costa, C. A. (2020). Enhancing knowledge construction processes within multicriteria decision analysis: The collaborative value modelling framework. *Omega*, 94, Article 102047. <https://doi.org/10.1016/j.omega.2019.03.005>
- von Winterfeldt, D., & Edwards, W. (1986). *Decision analysis and behavioral research*. Cambridge Univ. Press.
- White, L. (2006). Evaluating problem-structuring methods: Developing an approach to show the value and effectiveness of PSMs. *Journal of the Operational Research Society*, 57(7), 842–855. <https://doi.org/10.1057/palgrave.jors.2602149>