Gamified environmental multi-criteria decision analysis: information on objectives and range insensitivity bias

Alice H. Aubert[a],* (iD), Judit Lienert[a] (iD) and Bettina von Helversen[b,c] (iD)

[a]*Eawag: Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland*
[b]*Department of Psychology, University of Bremen, Bremen, Germany*
[c]*Department of Psychology, University of Zurich, Zurich, Switzerland*
E-mail: alice.aubert@eawag.ch [Aubert]; judit.lienert@eawag.ch [Lienert];
b.helversen@uni-bremen.de [von Helversen]

Abstract

Multi-criteria decision analysis (MCDA) is well suited to address complex public policy problems but could benefit from new tools to involve many laypeople. Online information on specialized topics could be more engaging by including game elements. This paper reports an experiment that assessed a gamified interface to (1) inform laypeople about the objectives to consider in wastewater management decisions, (2) assist them in constructing range-based preferences, and (3) provide a positive experience. We measured the effects with (1) a knowledge pre- and posttest, (2) the elicited weights and a range sensitivity index, and (3) an experience questionnaire based on self-determination theory. Answers from 174 participants indicated that participants learnt about the objectives and constructed preferences in both the gamified and control treatments. However, in neither were weights sufficiently adjusted. Our gamification making the ranges salient did not help overcome this bias. Both treatments were experienced as neutral to positive, the gamified being more entertaining. We discuss implications: if gamification of tools for participatory decision-making is to be promoted, it requires further research. Range insensitivity remains an unresolved bias in MCDA.

Keywords: behavioral OR; wastewater; preference elicitation; learning; Decision Support System; citizen participation; evaluation process; self-determination theory; gamification

1. Introduction

1.1. Background and motivation

Need for participatory multi-criteria decision analysis (MCDA). Public policy and environmental decision problems are complex, with many stakeholders involved that may have very different

*Corresponding author.

interests and values (e.g., Cockerill et al., 2019; Meinard et al., 2021). Increasingly, there are calls from various fields, including the OR community, to engage affected citizens in such decision-making processes (Lourenco and Costa, 2007; White and Bourne, 2007; Insua and French, 2010; Johnson et al., 2018; Vieira et al., 2020).

MCDA is particularly suited to addressing decision problems with conflicting interests (Keeney and Raiffa, 1976; Belton and Stewart, 2002). However, involving the public presents a challenge to the current tools and practices in MCDA (Insua and French, 2010; French and Argyris, 2018; Johnson et al., 2018). First, the affected citizens are usually new to the topic and need to acquire knowledge about facts to make an informed decision. For instance, water infrastructure management is an environmental issue of the highest importance to citizens' well-being (Pahl-Wostl et al., 2003) and merits involving a broad range of citizens. Such issues require factual knowledge, if people are asked to make trade-offs between unusual objectives such as high removal of organic matter and high flexibility of infrastructure systems (Larsen et al., 2016). Affected citizens first need to understand these concepts. Second, these citizens most likely have no prior preferences and need to construct them (Lichtenstein and Slovic, 2006). Third, these citizens are usually unfamiliar with the MCDA process and need to understand it to provide reliable responses. Facilitators usually elicit responses in face-to-face interviews and group workshops and ensure that responses are consistent and reliable. Specifically, facilitators should verify that the answers reflect the respondents' preferences (Payne et al., 2006). Among other recommendations, facilitators could support stakeholders through a training task to help them understand the elicitation process, before eliciting the preferences relevant in their case (Anderson and Clemen, 2013). However, often, many citizens are affected, and face-to-face interviews and workshops are not suitable. Surveys can involve numerous citizens, but have been criticized in MCDA for their lack of interaction and validation checks by decision analysts (Marttunen and Hämäläinen, 2008). The present study developed and tested a tool designed to enable broad participatory and unassisted MCDA.

Possible solution: Gamification. Gamification as an interdisciplinary research field is growing rapidly. Gamification is commonly defined as the use of "elements of game design in non-game contexts, products, and services to motivate desired behaviors" (Deterding, 2012), or as "a process of enhancing a service with affordances for gameful experiences in order to support users' overall value creation" (Huotari and Hamari, 2017). Originating from information and communication technology (Koivisto and Hamari, 2019), gamification often aims to increase participation and performance in online computer interactions (e.g., Darejeh and Salim, 2016; Mekler et al., 2017; Subhash and Cudney, 2018), and independent use of computer interfaces by laypeople (e.g., Keusch and Zhang, 2015; Morschheuser et al., 2016). However, gamification is rarely rigorously tested and appears as potentially promising (Aubert et al., 2019; Koivisto and Hamari, 2019). In this study, building on behavioral OR (Franco and Hämäläinen, 2016; Franco et al., 2021) and psychology, we use gamification to address the challenges that MCDA faces in becoming more participatory, and test its potential benefits.

Previous test of gamification for MCDA. In a previous study, we developed a gamified prototype to elicit the relative importance given to objectives by many stakeholders (Aubert and Lienert, 2019). Despite some positive results, we concluded that the proposed gamification needed stepwise improvement. Among other suggestions, we proposed (1) reducing long and tedious reading about the definitions of the objectives. Moreover, we suggested (2) designing the gamification based on the basic psychological needs theory (BPNT) to enhance the user experience (see Section 1.2 and,

e.g., Ryan, 2017). We also recommended (3) improving the research design, for example, by better differentiating the gamified and control treatments. In this study, we address these three points.

Structure of the paper. In Section 1.2, we provide theoretical background, and define three research questions. We introduce the gamified tool for providing information about objectives in Section 1.3. In Section 2, we present the experimental design, participant recruitment, the measures used, and the data analysis. We report the results in Section 3. We finish by discussing these results in Section 4 and reflect on the limitations of the proposed gamification in Section 4.4 and on follow-up research development in Section 4.5.

1.2. Theoretical background and research questions

Factual learning for MCDA. The present work builds on value-focused thinking (Keeney, 1992). This means deciding based on preferences regarding objectives, which should be fulfilled, rather than on preferences regarding alternatives. Preferences are usually constructed based on the available information, and the construction often occurs during the elicitation process (Lichtenstein and Slovic, 2006). Constructing preferences during elicitation is especially likely when there is no prior knowledge on the topic and no prior knowledge can be recalled (Hoeffler et al., 2013). Thus, preference construction requires factual learning: collecting and processing information about the objectives senses and meaning, their attributes and the units used to measure them, the ranges, and the associated impact of the attribute levels (Payne et al., 2006).

We developed the gamified release of information on objectives (Section 1.3) to increase factual learning. We focused on informing laypeople about the various objectives in sustainable wastewater management decisions. We used information from real decision cases in rural Switzerland, where individual households could change from current centralized to decentralized wastewater management systems (Beutler and Lienert, 2020). The best way to transmit information is still unclear. Literature on gamification for learning initially had high expectations and reported positive results (e.g., Plass et al., 2015). By raising affective and cognitive engagement (Dede, 2009; Connolly et al., 2012; Ramirez and Squire, 2014), and/or increasing the satisfaction of basic psychological needs (Ryan and Deci, 2017b), gamified learning increase learning performances. However, recent reviews are critical (Dichev and Dicheva, 2017; van Roy and Zaman, 2017). Evidence of long-term benefits of gamification are insufficient, and practice has outpaced research, leading to limited understanding of mechanisms and methods. Others claim that game-based learning fails due to inappropriate design that takes insufficient account of context and user characteristics (van Roy and Zaman, 2017). This leads to our first research question:

(RQ1) Does the proposed gamification of information on objectives facilitate factual learning about the objectives?

Preference construction based on ranges. During the process of factual learning, preference construction manifests itself by the creation of a preference if no prior preference is available, a change in preferences, or the consolidation of a prior preference (Belton and Elder, 1994; Aubert and Lienert, 2019). Preference construction can be subject to biases, which often stem from the use of heuristics (Morton and Fasolo, 2009; Montibeller and von Winterfeldt, 2015). Methods such as MCDA aim to promote rational decision-making by avoiding heuristics (Eisenführ et al., 2010).

Thus, training and careful use of such methods should mitigate the most severe biases and the use of heuristics. When eliciting weights, good practice requires to ask consistency check questions (Hobbs and Meier, 1994), emphasize the attributes ranges to counteract the range insensitivity bias (Gabrielli and von Winterfeldt, 1978; von Nitzsch and Weber, 1993; Eisenführ et al., 2010; Montibeller and von Winterfeldt, 2015), and use symmetrical hierarchies of objectives (Jacobi and Hobbs, 2007; Marttunen et al., 2018).

The present work contributes to behavioral OR (Franco and Hämäläinen, 2016; Franco et al., 2021). It focuses on local preferences, which are based on the attribute range in the specific context (Goldstein, 1990; van Ittersum and Pennings, 2012), rather than on general attitudes toward the objectives, also referred to as global preferences. In this study, local preferences are captured as weights given to objectives. According to the value-comparison hypothesis (Fischer, 1995), local weights for objectives are best elicited with methods that emphasize cross-attribute comparisons, such as trade-off and swing methods (e.g., Eisenführ et al., 2010).

For instance, when deciding about a job, a decision-maker might state that high salary is the most important objective. This is a general statement, based on a global scale from very low to very high salary. However, consider a specific decision where (1) the range of variation in salary is marginal, that is, the alternatives to choose from offer similar salaries, and (2) the range of variation of a second criterion unimportant on the global scale varies greatly on the local scale. For example, the number of holiday days per year could be four times larger for one alternative than for the others. A decision-maker will be asked to express trade-offs between "high salary," which varies marginally, against "high number of holiday days," which varies greatly. Considering the local scale of this specific decision, the decision-maker should give little importance to "high salary," because it does not discriminate well between the alternatives relative to "high number of holiday days," which does. The trade-off and swing weight elicitation methods emphasize such comparisons between attribute levels, and should be range sensitive. The value difference on a narrow attribute range is larger than that on a broad attribute range if the value function is renormalized. As consequence, the weights should be lower on a narrow attribute range than on a broad range.

However, experiments report that the weights adjustment for manipulated ranges of attribute is lower than the normatively assessed adjustment (Gabrielli and von Winterfeldt, 1978; von Nitzsch and Weber, 1993; Fischer, 1995). The range sensitivity principle is so generally accepted in the decision analysis literature (see, e.g., textbook Eisenführ et al., 2010, p.151-154), that to our knowledge very few recent follow-up experiments or analyses have been conducted, despite great concern about decision-makers' insufficient adjustment of weights to ranges (Morton and Fasolo, 2009; Montibeller and von Winterfeldt, 2015). Moreover, these early experiments used simplified decision problems such as job, flat, or car selection with four attributes or fewer. In our experiment, we used a complex real-world decision on wastewater management to investigate range sensitivity. Our second research question reads as follows:

(RQ2) Does informing participants on objectives and making attribute ranges especially explicit facilitate preference construction based on the provided range of the attribute? Specifically, do we observe the expected change in weights when the range is manipulated?

Gamification to improve the experience. User experience is central to interfaces evaluation. We used the self-determination theory (SDT), in particular the basic psychological needs theory

(BNPT), to investigate user experience of the gamified information on objectives, as is often done in the literature on gamification (Seaborn and Fels, 2015; Mekler et al., 2017; Nacke and Deterding, 2017; van Roy and Zaman, 2017; Ryan and Deci, 2017a). The original proposition reads, "there are three basic psychological needs, the satisfaction of which is essential to (…) integrity, and well-being. These are the needs for autonomy, competence, and relatedness. (…) Need frustration is associated with impoverished functioning." (Ryan and Deci, 2017b, Proposition 1a, p.242). Additionally, these authors propose that "psychological need satisfaction and frustrations vary within persons over time, contexts, and social interactions. Any factor or event that produces variations in need satisfaction or need frustration will also produce variations in wellness, and this principle extends from highly aggregated levels of analysis down to moment-to-moment or situation-to-situation variations in functioning." (Ryan and Deci, 2017b, Proposition 1b, p.243). Other findings from qualitative approaches, close to the grounded theory, align with the BPNT (Heimann and Roepstorff, 2018). In this study, we evaluate user experience by exploring whether gamification satisfies basic needs.

Previous studies suggest that gamification fulfills three fundamental human needs: autonomy, competence, and relatedness (e.g., Nacke and Deterding, 2017, and references therein). Autonomy represents the need for volition, to decide voluntarily for oneself. Providing choices is common to manipulate autonomy (Sheldon and Filak, 2008; Przybylski et al., 2010; Ryan and Rigby, 2019). Competence represents the need to achieve something, to be successful and effective. It can be manipulated by providing clear goals and unlocking the next difficulty level when easier levels are achieved (Ryan and Rigby, 2019). Relatedness represents the need to connect with others in mutual exchange. In single-player video games, nonplayer characters are a major feature influencing relatedness (Ryan and Rigby, 2019). They support the player and enable some "exchange with others." We designed our gamification to satisfy these three needs (Section 1.3), and in turn improve the user experience. Our third research question is:

(RQ3) Does the proposed gamified interface to inform laypeople on objectives provide a positive user experience? Do the constructs of the self-determination theory explain this effect?

To answer the three research questions, we carried out an experiment. We tested a gamified interface to inform laypeople on objectives, based on a real-world decision about sustainable wastewater management. We also varied some of the range information provided. The experimental design is fully described in Section 2.

1.3. Developing a gamified tool to inform laypeople on objectives

Here, we report on the gamification design following the method by Morschheuser et al. (2017). This method includes seven phases: project objectives (I); analysis of context and users (II); ideation (III); design of prototype(s) (IV); implementation of design of prototype (V); evaluation (VI); and monitoring (VII). The present paper describes the first six phases, focusing on the evaluation. Monitoring is beyond the scope of our work.

(I) Project objective. First, we identified the aim of the gamified tool as providing information on objectives in such a way that it makes the attribute range explicit and salient. We expect that this will increase range awareness and help overcome the range insensitivity bias. The tool should

inform on the objectives to consider, in our example when deciding about wastewater management, following the current best practices (Payne et al., 2006; Anderson and Clemen, 2013). Specifically, the gamified tool should provide each objective's name, its definition, its measure or attribute, the unit in which it is measured, the best, worst, and status quo levels of each attribute, and the expected impact of these attribute levels. It should draw attention to the ranges, that is, the best and worst possible levels of each attribute. In other words, the gamified tool should provide the facts needed to understand the decision at stake and thus facilitate factual learning. In turn, this should help laypeople to construct range-sensitive preferences on the relative importance of the objectives. Moreover, the interface should provide an engaging user experience. Ideally, the gamified tool should be flexible enough to allow case-specific variations, for instance, in the objectives to consider or in the ranges of the attributes. To be accessible to many, it should be an online interface not requiring software download. An online interface would also allow centralized data collection from distributed respondents. As final aim, it should be possible to extend and merge this gamified tool with a prototype that we have developed to elicit weights assigned to objectives (Aubert and Lienert, 2019). Given our objectives, we decided to develop a "soft gamification" (Bailey et al., 2015). Soft gamification means that we included game elements in our informative interface, as opposed to a fully fledged game that includes information.

(II) Context and users. The target users are citizens affected by decisions about wastewater management in rural Switzerland. Thus, the target users are German-speaking adults and young adults with at least basic education. They will use the gamified tool online without assistance from an experienced decision analyst.

(III) Ideation. First, we brainstormed a storyline. The storyline should introduce a challenge motivating users to learn about wastewater management, hereby enabling them to understand the objectives. To develop the story, we considered the feedback from game designers concerning our first gamified prototype (Aubert and Lienert, 2019). They appreciated the initial storyline, but made several suggestions for improving it. In a nutshell, the game designers suggested that the story should become more immersive by connecting more strongly to the wastewater topic and local decision context. Second, we consulted game element lists (Duke, 1980; Hamari et al., 2014), and added those elements expected to positively influence the constructs of basic psychological needs theory (Table 1) that had not yet been included. We based this brainstorming phase on value-focused thinking (Keeney, 1992). This means that we developed options for telling an interesting story about wastewater management that would enable us to reach the objectives identified by game designers and the theories used to explain the mechanism of gamification.

(IV) Prototype development and pretests. In the early 2018, we twice pretested prototype versions drawn on paper at the Zurich University of the Arts (https://gamedesign.zhdk.ch/, retrieved on 11 October 2019) with game design lecturers. Each pretest led to improvements, resulting in a single, improved prototype. Thereafter, two decision analysts with experience in wastewater management tested the improved paper prototype. This enabled further development and sparked discussion to select the objectives, based on a Swiss case study (Beutler and Lienert, 2020). An internal workshop was organized with three decision analysts involved in real-case MCDA applications for decisions about wastewater management in Switzerland. We identified 10 objectives (Haag et al., 2019; Beutler and Lienert, 2020), and organized them in a two-level hierarchy of objectives. High removal of organic matter and high removal of micropollutants defined the higher level objective of high environmental protection. High recovery of nutrients, low water consumption, and low net energy

Table 1
Summary of the gamification concept. Each page (1 to 5) brings additional information on objectives to learn facts about wastewater management and to construct context-specific preferences for later use in MCDA (bold lines). Each user action corresponds to a game element, which is expected to influence one of the three constructs of the basic psychological needs theory (BPNT)

User's action	Game element	Expected BPNT construct
Page 1 – Context introduction, presentation of the problem		
Read / listen to Mr. Akles	Nonplayer character	Relatedness
Chose representation of self	Avatar	Relatedness
Page 2 – Definition of the 10 objectives		
Explore / discover virtual world	Map / move	Autonomy
Collect icons	Progress	Competence
Take short cuts, answer phone call, receive weather forecast alert	Random event	Autonomy
Page 3 – Definition of attributes and their units		
Match icons with measurement units	Clear goals / matching	Competence
Read / listen to encouragement from Mr. Akles	Feedback	Relatedness
Drink coffee offered by Mr. Akles	Random event	Autonomy
Page 4 – Presentation of best, worst (i.e., range), and status quo levels of attributes		
Explore / discover virtual world	Map / move	Autonomy
Meet a black raven or a white dove	Random event	Autonomy
Page 5 – Link to weight elicitation survey		
Receive a gift	Reward	Competence

consumption defined high sustainable resource use. High health protection, high attractiveness, and low time demand for the end-users defined high societal well-being. Low cost and high flexibility defined high economic performance. We wrote specification documents describing the design (available upon request). These included a technical document (eight pages) specifying following features: supporting technology and exploitation system, structure of the information on objectives and alternatives provided to the player, game elements, language, data that needs to be recorded, access, license, and confidentiality terms. Specifically, the description of game elements included an overview of the five chapters of the story, a description of the characters in the story, the feedback provided by the interface and the progression display, and the random events. An appendix to the technical document (19 pages) specified an exemplary path through the interface with pictures of the paper prototype, and text for the instructions and dialogues.

(V) Prototype implementation. The company Opinion Games GmbH (Brugg, Switzerland: https://opiniongames.ch/, retrieved on 15 August 2019) won the competitive call to implement the concept, described in the specification documents, and deployed it for web-based use. The gamification consisted of a narrative frame composed of five chapters, each being a web page (Figure 1, screenshots and descriptions of each chapter, see Supplementary Information SI 1). The overall mission is to discover the challenges Mr. Akles, a wastewater engineer, faces. The challenges correspond to the objectives for MCDA. The designed gamification, based on the BPNT, released the information gradually. It included a chapter dedicated to the range of attributes. Autonomy game elements (action choice and random events) should make the users feel volition. Competence game elements consisted of small clearly defined tasks to achieve, and gradual information delivery. Hereby,
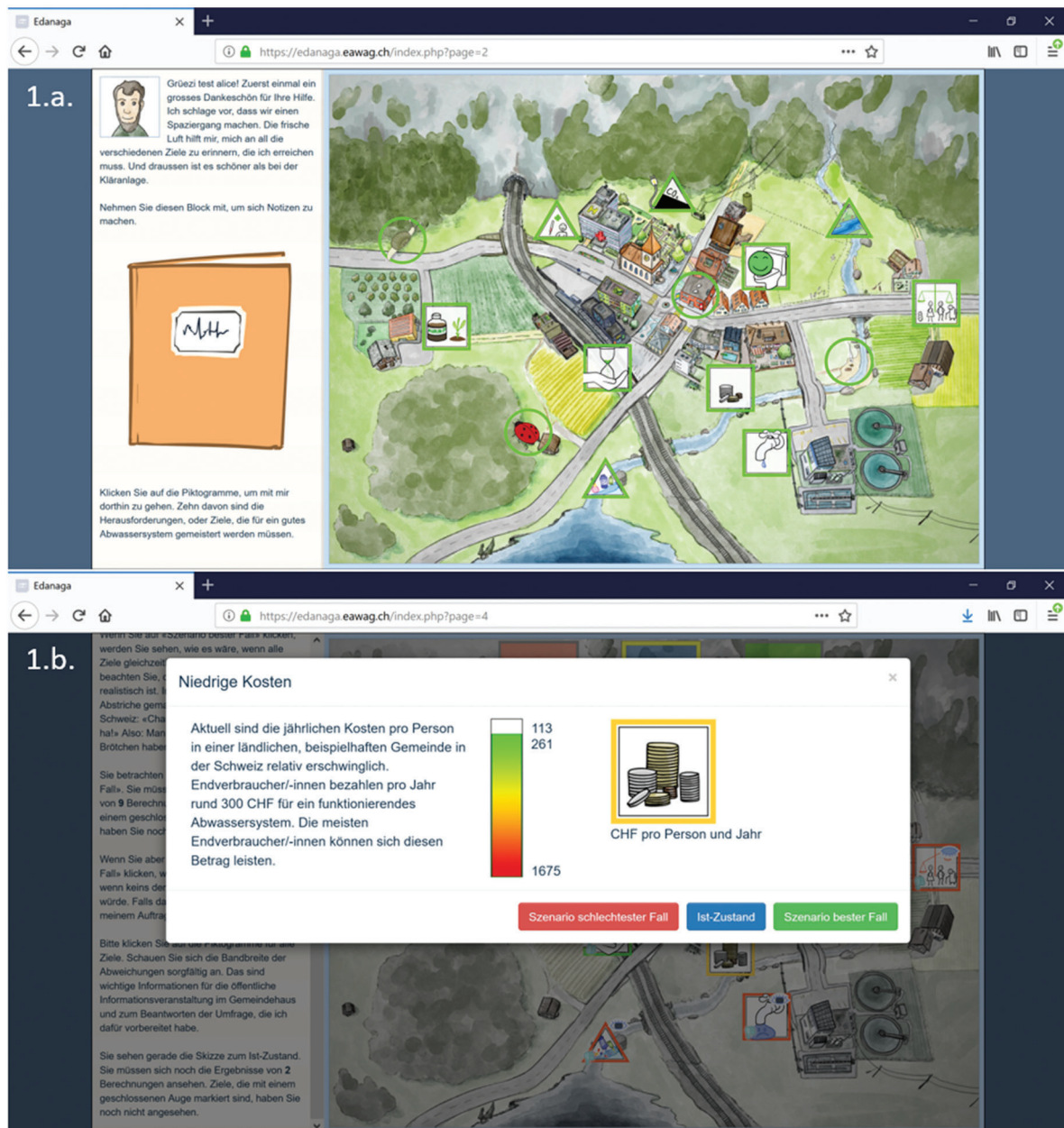
Fig. 1. Screenshots of the gamified information on objectives for a wastewater management decision in rural Switzerland. (a) Page 2: discovering the objective (name, icon, and definition). (b) Page 4: pop-up window depicting an attribute level in the status quo, the current state in the decision case; that is, the fulfilment bar for "low costs," the text (in German) describing the attribute level and its expected impact, the attributes' unit (e.g., CHF per person and year), and the numerical values for the worst (red, lowest number), and best states (green, number at the top). Clicking on the specific red ("worst possible state") or green ("best possible state") buttons changes the fulfilment bar and the descriptive text.
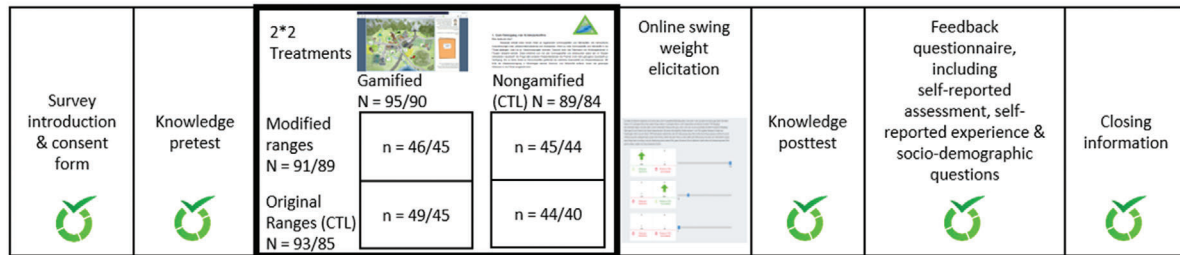
Fig. 2. Experiment flow of the 2 × 2 design, including sample sizes after data cleaning. For RQ1 and RQ3, the total sample size was 184 participants, and for RQ2 174. First numbers indicate the number of participants per group for $N = 184$, second numbers for $N = 174$. CTL: control treatment. Participants started with a pretest (programmed in LimeSurvey), and were then automatically directed to one of the four treatments, thereafter to a weight elicitation interface, and finally to a survey (programmed in LimeSurvey).

the aim was that the users should feel progress and could realize achievements. Relatedness game elements consisted of interacting with a guiding nonplayer character, Mr. Akles. Table 1 also summarizes the users' actions and related game elements.

(VI) Evaluation. The focus of this paper is on the evaluation of the gamified interface prototype. We provide method details concerning evaluation in the Introduction (Section 1.2), in the Methods (Section 2), and the results in Section 3.

2. Data and Methods

2.1. Experimental design and procedure

We evaluated our gamified interface quantitatively and qualitatively. First, we compared the gamification (Fig. 1) with a control treatment. The control consisted of a pdf file containing the identical information on the 10 objectives. Each objective description fitted on an A4 page (example screenshot, see SI 2). The experiment was conducted in a controlled environment in a lab at the Institute of Psychology at the University of Zurich. Participants completed the survey using the university computers seated in individual cubicles.

We designed a 2 × 2 between-subject experiment. The varying factors were (1) the information mean (gamified vs. nongamified control) and (2) the attribute ranges (manipulated vs. original control), leading to four treatments in total (Fig. 2).

When arriving, participants received an information sheet about the experiment and signed a consent form. Thereafter, they answered a pretest on the computer about their knowledge of wastewater management (SI 3). One of four treatments followed (Fig. 2). Next, we elicited participants' preferences, specifically the weights, using the swing method (von Winterfeldt and Edwards, 1986; Eisenführ et al., 2010). Its procedure makes attribute ranges explicit (description see SI 4, with screenshots). Therefore, swing elicitation is recommended to limit the range insensitivity bias (Montibeller and von Winterfeldt, 2015). How to elicit swing weights online is discussed elsewhere (e.g., Aubert et al., 2020). Finally, the knowledge test was repeated (SI 3), and participants responded to a feedback questionnaire, designed to measure the autonomy, competence, relatedness, perceived learning, and experience (SI 5).

Second, we qualitatively evaluated the gamification orally and in writing to strengthen our understanding of the quantitative results. The written qualitative data were answers to open text boxes in the feedback questionnaire. The oral qualitative data were collected during a test user workshop with game design students from Zurich University of the Arts, which was facilitated by the first author. After introducing the BPNT, participants signed an informed consent form and individually tested the gamified interface. Thereafter, they individually assessed it based on three tasks (SI 6). First, they wrote down two words that spontaneously came to their mind to describe their experience. Second, they rated their overall experience on scales from 1 to 5 (selection of questions from the experiment). After a break, we discussed their answers in an audio-recorded session.

Finally, we tested the gamified interface in a real case study (without a control). We assumed that the real-case respondents, whom the decision substantially affects, would engage in the survey more than participants from the experiment would. A Swiss rural municipality was deciding about the future of their wastewater management following a structured MCDA process in facilitated workshops (Beutler and Lienert, 2020). The local decision-makers wanted to include the citizens' opinions, since some alternatives directly affect them. For instance, novel decentralized wastewater technologies could be installed in their houses (Larsen et al., 2016). The local decision-makers welcomed the gamified information on objectives and the weight elicitation survey. We included a shortened version of the feedback questionnaire. The municipal administration invited the citizens to participate between August and September 2019. A lottery incentivized the respondents. Its total value was CHF 1000, which was split among respondents who registered as vouchers for one of four stores.

2.2. Sample definition and recruitment

We randomly distributed the four treatments among participants and balanced for gender. In addition, to limit the potential confounding effect of age due to the digital divide between generations (Koivisto and Hamari, 2014), we focused on the age group of 18–30 years. This age group is (1) part of the target group, being citizens of voting age, (2) used to digital interfaces, and consequently (3) familiar with their design and usability (Cunliffe, 2000). Thus, they are unlikely to drop out of the survey for lack of digital knowledge and can provide critical feedback, both of which are important for informing future improvements.

We advertised the experiment on the internal pin board at Eawag, on the student virtual marketplace (marktplatz.uzhalumni.ch/de, retrieved on 24 January 2020) of the ETH Zurich and the University of Zurich (UZH), on the test person server of the UZH (www.psychologie.uzh.ch/probandenserver/public/, retrieved on 24 January 2020), on mini-jobs.ch, with flyers in various ETH and UZH buildings, and the ad was further shared on social media and per e-mails. The invitation was in German. Each participant received CHF 30.

We determined the sample size by *a priori* statistical power analysis using G*Power 3.1.9.2 (Faul et al. 2009), with the results reported in Ryan (2017). We targeted 200 participants to be equally distributed across our four treatments. We collected data from January to mid-March 2019. The project underwent ethical review by the Eawag directorate and was approved of; it classified as a minimal risk project involving human subjects.

2.3. Measures

For each research question, we developed measures based on the literature (Table 2). Some measures are quantitative (performance test), while others are subjective self-reported evaluations by each participant. For the sake of conciseness, we only develop the Range Sensitivity Index hereafter. The specifics of all measures are in the SI.

To measure preference construction, we elicited swing weights given to the 10 objectives (full procedure described in SI 4, with screenshots). First, participants rank-ordered hypothetical scenarios with only one objective at its best attribute level. Second, participants rated the ordered hypothetical scenarios relatively to one another, assigning points between 100 (for the preferred scenario that was ranked first) and 0 (for the worst-case scenario, where all objectives were set to their worst level as default). We explored possible effects of range sensitivity by informing on objectives. We made the attribute ranges explicit.

To investigate RQ2, the ranges of attribute was a varying factor. Previous experiments halved or doubled the ranges (von Nitzsch and Weber, 1993; Fischer, 1995). Gabrielli and von Winterfeldt (1978) called for a strong manipulation of attribute ranges, as they observed that they may not have manipulated the ranges sufficiently in their experiments. We modified two of the original ranges of our case. We chose two attributes that (1) were expressed in the same readily understood unit, the percentage, and (2) could vary in opposite directions. The objective of high removal of organic matter (Orgm) was made relatively unimportant by reducing the range of its attribute: the original worst level of only 37% removal of organic matter was increased to 82%, so the original range of 56% was reduced to 11%. The objective of high flexibility (Flex) was made relatively important by increasing the range of its attribute: the original worst level of 37% flexibility was further decreased to 1%, so the original range of 51% increased to 87%.

We decided for a between-participant experiment for two reasons (Fischer, 1995). First, it prevents subjects from directly comparing attribute ranges and adjusting weights accordingly. Second, it is closer to real-world conditions: when making a decision, there is usually only a single local context. Thus, we compared the weight distributions (mean, median, and standard deviation) of the two objectives with manipulated attribute ranges to the weight distributions with the original ranges. We calculated the range sensitivity index (RSI) developed by von Nitzsch and Weber (1993) across participants, with the reference objective of low cost (Cost), which is easy to understand. We also calculated the RSI with a second reference objective of low water consumption (Wat), which statistically had the same average weight in both conditions (8.5 (SD = 3.1) and 8.2 (SD = 3.3), respectively; based on the results of the experiment). The RSI consists of the percentage of range difference that the participants considered. It is calculated with M, the expected change of weight resulting from the normative analysis of attribute range effects, in our case with a linear value function, and M^{eli}, the elicited or empirical change of weight (von Nitzsch and Weber, 1993; Fischer, 1995; Eisenführ et al., 2010), as in Equation (1):

$$RSI = \frac{M^{eli} - 1}{M - 1}. \tag{1}$$

We averaged the weights between participants and scaled the weights of Orgm (or Flex) and Wat (or Cost) so that (1) the ratio between Orgm (or Flex) and Wat (or Cost) stayed constant; and (2)

Table 2
Measures used for the three research questions

Research questions	Measure type	Measures	Ref. & SI
1. Does the proposed gamification of information on objectives facilitate factual learning about the objectives?	Quantitative (performance test)	Difference in Knowledge Score: repetition of the same 10 questions (one per objective) before (initial) and after (final) informing on objectives; varies from -10 to 10; positive ΔKS indicates factual learning; pretested (previous studies + dedicated data collection with 52 participants);	Lienert et al. (2016) SI 3 SI 8a
	Self-reported	Three feedback questions (5-point Likert scales: 1 "no factual learning at all" to 5 "a lot of factual learning")	SI 5a
2. Does informing participants on objectives and making attribute ranges especially explicit facilitate preference construction based on the provided range of the attribute? Specifically, do we observe the expected change in weights when the range is manipulated?	Quantitative (performance test)	• Range Sensitivity Index (RSI); based on elicited weights (between-participants), 0 means no adjustment of the weights, 1 weight adjustment to the new range as normatively expected	von Nitzsch and Weber (1993), Fischer (1995)
		• Perceived range sensitivity: four feedback questions (two per manipulated objective); varies from 0 (all answers were wrong) to 4 (all answers correct)	SI 5b
	Self-reported	• Two feedback questions (5-point Likert scales: 1 "ranges not considered at all" to 5 "ranges very much considered")	SI 5c
		• Two feedback questions (5-point Likert scales: 1 "no preference construction at all/answers do not represent opinion at all" to 5 "learnt a lot about preference/answers exactly represent opinion")	SI 5a
3. Does the proposed gamified interface to inform laypeople on objectives provide a positive experience? Do the constructs of the basic psychological needs theory explain this effect?	Self-reported	Overall experience: eight feedback questions: entertainment, fun, no boredom, no irritation, no worry, contentment, recommendation, and worthwhile use of time (5-point Likert scales: 1 "not at all" to 5 "very much")	Aubert and Lienert (2019), Harms et al. (2014) SI 5a

Continued

Table 2
(Continued)

Research questions	Measure type	Measures	Ref. & SI
	Self-reported	Basic psychological needs (5-point Likert scales: 1 "not at all" to 5 "very much"): • Competence: six feedback questions (competence in factual learning (3), in preference construction (2), and perceiving progress (1)) • Autonomy: five feedback questions (focused on personal interest (2) and on the perceived availability of choices (3)) • Relatedness: three feedback questions (two about the participant's relationship to the societal issue of wastewater and one about the helpfulness of the information provided)	Deci and Ryan (2000), Gagné (2003), Peng et al. (2012), Przybylski et al. (2010), Ryan and Deci (2017a), Sheldon and Filak (2008), Tamborini et al. (2010), Xi and Hamari (2019), SI 5a

the sum of the weights of Orgm (or Flex) and Wat (or Cost) equaled 1. Thereafter, we used the largest range as the global value model and calculated M, the expected adjustment factor of the weight for the narrow range conditions as in Equation (2):

$$M = \Delta v_r \left(B'_r \right), \tag{2}$$

where v_r represents the linear value function for Orgm (or Flex) on the global attribute range, and B'_r the local attribute range. We calculated M^{eli}, the elicited (or empirical) adjustment factor of the weight as in Equation (3):

$$M^{eli} = \frac{w_{local}^{manip} / w_{local}^{ref}}{w_{global}^{manip} / w_{global}^{ref}}, \tag{3}$$

with the scaled average of the elicited weights as w_{local}^{manip} for the objective Orgm (or Flex) in the narrower attribute range treatment, w_{local}^{ref} for the objective Wat (or Cost) in the narrower attribute range treatment, w_{global}^{manip} for the objective Orgm (or Flex) in the larger attribute range treatment, and w_{global}^{ref} for the objective Wat (or Cost) in the larger attribute range treatment. According to the range sensitivity principle, the RSI should be 1 (elicited weight normatively adjusted). If the weights are range insensitive, the RSI equals 0.

2.4. Data analysis

We analyzed the data using R project for statistical computing (R Development Core Team, 2020). Given the sample size (>50), we evaluated the differences between groups using Student t-tests. Effect sizes were calculated with the *effsize* package. Further details are given in Section 3.

We removed 20 participants from the entire analysis due to their responses to a filter question (SI 5d). The filter question asked whether participants had encountered a nonplayer character. Obviously, participants receiving the gamified information should answer yes, the others no. Six participants receiving the gamified treatment answered that they had not seen a character. Fourteen participants receiving the control answered that they had seen a character. We discarded all 20, because we doubted the reliability of their answers.

3. Results

3.1. Participants of the experiment

In total, 204 people participated in our experiment. We analyzed data from 184 participants, of whom 132 were female and 51 male. One person did not disclose gender. The mean age of participants was 23.5 years ($SD = 3.16$). The majority (75.54%) had a university degree. Most participants had little or no previous knowledge of wastewater: 97.28% answered knowing "nothing at all," "rather little," or "a little" about wastewater before the experiment. The sample was split in two groups, gamified ($N = 95$) and nongamified ($N = 89$). The gender ratio was equal in both groups (see SI 7 for a detailed sample description).

3.2. Factual learning (RQ1)

Overall, the participants' delta knowledge score (ΔKS, difference between the initial and final scores) was positive (Table 3): they learnt about the objectives after receiving information. The mean initial and final scores were statistically significantly different (paired t-test; $t(183) = -23.08$, $p < 0.001$, $d = -1.97$). There was no difference in the scores between the nongamified and the gamified treatments, suggesting that a similar amount of factual learning occurred in both treatments (two sample t-test; $t(182) = -0.74$, $p = 0.462$, $d = -0.11$).

Note, we tested the knowledge test with 52 additional participants that completed the knowledge test twice, six days apart (median) but did not receive information on objectives. This test showed that (1) overall, without information on objectives, there were no statistically significant differences between the initial and final scores; (2) within participants, the initial and final score correlated, indicating good test-retest reliability, $r = 0.75$; and (3) the ΔKS from our informed participants, those participating in the actual experiment, were statistically significantly higher than the ΔKS from uninformed participants (see SI 8a, for results of the statistical test). Thus, we are confident that our interventions facilitated factual learning.

There was also no statistically significant difference between the treatments in the self-reported factual learning ratings on 5-point Likert scales (SI 5a). Participants reported having "rather learnt" about wastewater management (gamified: mean $\mu = 3.88$, $SD = 0.68$; nongamified $\mu = 3.93$, $SD = 0.77$). In addition, the 20 citizens from the Swiss case study (Beutler and Lienert, 2020) obtained final knowledge scores in the same order of magnitude ($\mu = 6.4$, $SD = 2.1$, median $Mdn = 7.0$).

Table 3

Knowledge scores (KS) for the pretest before receiving information on objectives (initial) and the posttest after receiving information (final). KS varied between 0 and 10. ΔKS = final KS – initial KS; thus, $-10 < \Delta$KS < 10. μ = mean; SD = standard deviation; Mdn = Median

	Total ($N = 184$)		Gamified ($N = 95$)		Nongamified ($N = 89$)	
	μ (SD)	Mdn	μ (SD)	Mdn	μ (SD)	Mdn
Initial KS	4.89 (1.27)	5.00	4.77 (1.26)	5.00	5.02 (1.28)	5.00
Final KS	7.58 (1.33)	8.00	7.51 (1.27)	7.50	7.66 (1.39)	8.00
ΔKS	2.69 (1.58)	3.00	2.74 (1.64)	3.00	2.65 (1.52)	3.00

Table 4

Mean weights (μ), standard deviations (SD), and median weights (Mdn) given to the objectives of high removal of organic matter (Orgm) and high flexibility (Flex), for the original (Origin.) and manipulated (Manip.) ranges of attribute

		Total ($N = 174$)			Gamified ($N = 90$)			Nongamified ($N = 84$)		
		N	μ (SD)	Mdn	N	μ (SD)	Mdn	N	μ (SD)	Mdn
Orgm	Origin.	85	14.10 (4.74)	14.10	45	15.10 (4.77)	14.47	40	14.73 (4.76)	13.83
	Manip.	89	13.13 (5.90)	12.50	45	13.68 (5.67)	12.87	44	12.58 (6.13)	11.92
Flex	Origin.	85	9.40 (4.34)	9.10	45	8.63 (3.46)	8.74	40	10.26 (5.06)	9.62
	Manip.	89	9.95 (4.33)	10.48	45	10.39 (4.91)	10.56	44	9.49 (3.64)	10.30

3.3. Preference construction (RQ2)

For RQ2, we discarded another 10 answers ($N = 174$). These 10 participants wrongly entered their identification number in the swing weight elicitation survey, which prevented us from linking their weights to the rest of the data. We compared the effect of the range manipulation on the weights using *t*-tests for independent samples.

By decreasing the range of the objective "high removal of organic matter" (Orgm) in the manipulation, we expected Orgm to receive a lower weight in the manipulated treatment than in the control. This was statistically significantly verified for the total sample, albeit with a small effect size (Table 4; SI 8b). By increasing the range for the objective "high flexibility" (Flex) in the manipulation, Flex should receive a higher weight in the manipulated treatment than in the control. Over the total sample, the weight given to Flex in the manipulated treatment was higher than in the control; however, the difference was not statistically significant (Table 4; SI 8b). The range manipulation seemed to have a small effect on the weights but was mostly not significant, once even in a counterintuitive direction (mean weight for Flex in the nongamified group), and not systematically in favor of gamification, despite our design that made attribute ranges explicit.

The range sensitivity indices (RSIs) varied from negative values to 0.63 at best (Table 5). The RSI is the ratio between the expected and the elicited weights (see Methods, and von Nitzsch and Weber, 1993). Negative RSIs, not reported so far in the literature, imply that the elicited weights lead to $M^{eli} > 1$. This means that the ratio of the elicited weights for the narrow attribute interval was higher than the ratio of the elicited weights with a larger attribute interval. We observed

Table 5

Expected (M) and elicited (M^{eli}) weights (mean (standard deviation)), and range sensitivity index (RSI) for the objectives "high removal of organic matter" (Orgm) and "high flexibility" (Flex), in reference to the objectives "low water consumption" (Wat) and "low cost" (Cost). See Methods for explanation of RSI

| | N | Orgm (vs. Wat) | | | Orgm (vs. Cost) | | |
		M	M^{eli}	RSI	M	M^{eli}	RSI
Total	174	0.26 (0.20)	0.61 (0.91)	0.11	0.23 (0.20)	0.56 (0.83)	0.21
Gamified	90	0.25 (0.20)	0.63 (1.0)	0.0	0.23 (0.20)	0.57 (0.88)	0.15
Nongami.	84	0.27 (0.20)	0.59 (0.81)	0.24	0.23 (0.20)	0.55 (0.79)	0.27

| | N | Flex (vs. Wat) | | | Flex (vs. Cost) | | |
		M	M^{eli}	RSI	M	M^{eli}	RSI
Total	174	0.41 (0.59)	0.52 (0.91)	0.21	0.36 (0.59)	0.49 (1.0)	0.0
Gamified	90	0.43 (0.59)	0.49 (0.74)	0.63	0.37 (0.59)	0.46 (0.86)	0.34
Non-gami.	84	0.40 (0.59)	0.56 (1.1)	−0.35	0.35 (0.59)	0.52 (1.2)	−0.42

negative RSIs for high flexibility (Flex) in the nongamified group (consistent with the mean weight results (Table 4)). It indicates that the weights were adjusted to the range in the opposite direction from that suggested by the range sensitivity principle. We observed the highest positive RSI for the gamified group (Flex). In contrast, the nongamified group was more range sensitive than the gamified group for the objective "high removal of organic matter" (Orgm). In sum, the results of gamification on range sensitivity were equivocal: our gamification emphasizing the attribute ranges when informing participants on objectives did not systematically lead to the normatively expected weight adjustment.

When testing the perceived range sensitivity (feedback questions) about the percentage of improvement between the best and worst cases (see method Section 2.3.2), participants who received the nongamified treatment achieved a statistically significant better score ($\mu = 2.11$, $SD = 0.93$, $Mdn = 2.00$) than those who received the gamified treatment ($\mu = 1.52$, $SD = 1.04$), $Mdn = 1.00$; $t(172) = 3.90$, $p < 0.001$, $d = 0.59$). However, the self-reported range sensitivity on the 5-point Likert scales did not differ between the treatments: on average, participants reported being somewhat (3) to rather aware (4) of the best and worst levels (gamified: $\mu = 3.41$, $SD = 0.69$, $Mdn = 3.50$; nongamified: $\mu = 3.48$, $SD = 0.61$, $Mdn = 3.50$; SI 8b). Most participants reported learning about their own preferences in both treatments (gamified: $\mu = 3.51$, $SD = 0.74$, $Mdn = 4$; nongamified: $\mu = 3.44$, $SD = 0.67$, $Mdn = 3$; SI 8b). They also reported that those preferences reflect their opinions (gamified: $\mu = 4.35$; $SD = 0.66$, $Mdn = 4$; nongamified: $\mu = 4.28$, $SD = 0.64$, $Mdn = 4$; SI 8b).

3.4. Experience (RQ3)

We measured how users experienced the interface for informing about wastewater objectives with eight items (see SI 5a). We then compared this perceived experience in the gamified and nongamified treatments using independent sample t-tests and adjusting the significance level with Bonferroni corrections for multiple testing. The gamified group experienced the survey as

© 2022 The Authors.
International Transactions in Operational Research published by John Wiley & Sons Ltd on behalf of International Federation of Operational Research Societies

Table 6
Results for the measures of experience, for the gamified and nongamified (control) treatments, and statistical difference tests

	Gamified ($N = 95$)		Nongamified ($N = 89$)		Test: Two Sample t-test		
	μ (SD)	Mdn	μ (SD)	Mdn	t(df)	p-value (Bonferroni corrected)	Cohen's d
Entertainment	3.53 (0.90)	4	2.99 (0.95)	3	−3.95 (182)	<0.001 (0.001)	−0.58
Fun	3.42 (0.83)	3	3.28 (0.87)	3	−1.12 (182)	0.265 (1)	−0.17
No boredom	3.99 (0.88)	4	3.99 (0.95)	4	−0.01 (182)	0.996 (1)	−0.00
No irritation	4.16 (0.78)	4	4.27 (0.81)	4	0.96 (182)	0.340 (1)	0.14
No worry	3.81 (1.15)	4	3.78 (1.11)	4	−0.14 (182)	0.886 (1)	−0.02
Contentment	3.54 (0.87)	4	3.48 (0.87)	4	−0.42 (182)	0.676 (1)	−0.06
Worthwhile use of time	4.06 (0.76)	4	4.00 (0.78)	4	−0.56 (182)	0.578 (1)	−0.08
Recommendation	4.00 (0.91)	4	4.03 (0.82)	4	0.26 (182)	0.793 (1)	0.04

significantly more entertaining than the nongamified group (t(182) = –3.95, $p < 0.001$, d = –0.58, Table 6). No differences were found between the groups for any other of the single items (Table 6).

To our knowledge, no properly validated scales exist to measure the constructs of basic psychological needs when studying the effects and mechanisms of gamification. Therefore, as is common in game and gamification studies, we adapted the scales that we found in the literature (Sheldon and Filak, 2008; Przybylski et al., 2010; Tamborini et al., 2010; Peng et al., 2012; Ryan and Deci, 2017a; Xi and Hamari, 2019). We did not attempt to create a new scale, which would require testing for convergent and discriminant validity. The following result can be considered as a prestudy. Our scales to measure competence showed acceptable internal reliability ($\alpha = 0.74$), comparable to internal reliability levels reported in the literature (e.g., Sheldon and Filak 2008; Przybylski et al., 2010; Ryan, 2017). This indicated that the items consistently measured a single construct. The internal reliability for our scale for autonomy was also acceptable ($\alpha = 0.70$). However, the internal reliability for our scale for relatedness was not acceptable ($\alpha = 0.19$), indicating that the items did not measure a single construct. We thus refrained from analyzing its items on a single scale (i.e., aggregating all items) and report comparisons at the single-items level. This result suggests that two different subdimensions of relatedness may not be well captured on a single scale (discussed in Section 4.3). There were no statistically significant differences between the two treatments for any of the items measuring experienced competence, autonomy, or relatedness (Table 7). For all, answers were neutral to positive (3 and above).

Autonomy and competence were positively correlated with entertainment (for $N = 184$, $r = 0.40$, $r = 0.40$, SI 8c). These correlations were of the same order of magnitude for the gamified ($r = 0.38$, $r = 0.45$ for autonomy and competence, respectively) and the nongamified groups ($r = 0.41$ and $r = 0.36$). However, the trends that competence correlated more closely with entertainment in the gamified group than in the nongamified one and that autonomy correlated more closely with entertainment in the nongamified group than in the gamified one were not significant (SI 8c). Nevertheless, the need for autonomy and competence seemed to have been satisfied, which potentially enhanced the user experience.

Table 7
Mean (M), standard deviation (SD), and median (Mdn) of the constructs of the basic psychological needs theory for all participants (total), participants from the gamified treatment and from the nongamified treatment. Two-sample t-test to compare gamified and nongamified treatments. [a]Bonferroni corrected p-value

Construct	Cronbach's alpha	Total (N = 184) μ (SD)	Mdn	Gamified (N = 95) μ (SD)	Mdn	Nongamified (N = 89) μ (SD)	Mdn	Gamified versus Nongamified Two Sample t-test t	p-value	Df	D
Autonomy	0.70	3.50 (0.67)	3.60	3.55 (0.66)	3.80	3.45 (0.69)	3.40	−0.99	0.326	182	−0.15
Competence	0.74	3.93 0.46	3.92	3.95 0.44	3.83	3.92 0.48	4.00	−0.41	0.680	182	−0.06
forsociety1		3.63 (1.34)	4.00	3.66 (1.34)	4.00	3.58 (1.34)	4.00	−0.40	1[a]	182	−0.06
forsociety2 Relatedness		4.77 (0.50)	5.00	4.74 (0.55)	5.00	4.80 (0.43)	5.00	0.83	1[a]	182	0.12
helpful		4.21 (0.81)	4.00	4.24 (0.80)	4.00	4.18 (0.82)	4.00	−0.52	1[a]	182	−0.08

Finally, we checked for the relatedness with the nonplayer character (NPC) for the gamified group (SI 8d). Participants reported rather low relatedness to the NPC, with items receiving low scores between 2.16 for the creation of virtual relationship (worst 1, best 5) and 3.26 for how encouraging the NPC was.

3.5. Workshop, Swiss case study, and qualitative results

This section focuses on two additional data sources: the workshop with game design students and the Swiss case study. On 29 April 2019, 11 game design students and two lecturers took part in the workshop. The feedback on the general experience ranged from positive to negative (results of the first workshop task, SI 9a). The gamified information was perceived as informative, interesting, and a source of information. However, the amount and level of information was perceived differently: some found it was too complex, overwhelming, and demotivating (SI 9a), whereas others thought it was too simplified and not detailed enough. Participants discussed the game elements critically, both positively and negatively. For instance, one student warned that it was possible to click through without reading the text. The artwork was positively perceived by three participants, although one student reported mismatches between the icons and background styles. SI 9b presents comments about the nonplayer character Mr. Akles, who was intended to create relatedness. Many participants reported being annoyed by the automatic voice-over: they found that it sounded awkward and unnatural even though a professional actor had read the text. Four of the game design students qualified the gamification as bland (not spicy at all), boring, and monotonous, while positive comments included friendly and cute (SI 9a). Four participants asked for the target audience, doubting that the gamification was suitable for the citizens of a rural area. This last comment resonated with the qualitative feedback collected in the main experiment and the case study (SI 9b, SI 10b). Game design students suggested ways to improve the gamification (Sections 4.3 and 4.4).

In the Swiss municipality, 20 citizens completed both the weight elicitation and the shortened feedback surveys. Three of the respondents were decision-makers involved in the complete MCDA process and participated in decision-making workshops (Beutler and Lienert, 2020). The gamified information on objectives was perceived either positively or negatively (SI 10c). All items about the overall experience scored from neutral (3.1) to positive (4.5) (SI 8f).

4. Discussion

This paper focuses on using gamification for informing laypeople about the objectives important for wastewater management decisions in Switzerland. We tested whether a gamified interface highlighting the ranges enabled factual learning and range-sensitive construction of preferences, and whether it provided a positive user experience. We compared the interface with a nongamified control treatment in an experiment, and used the gamified version in a real decision case. In the following, we reflect on our three research questions (Sections 4.1–4.3). Since our study showed the limits of our proposed gamification, we then suggest ways on how to overcome and address the encountered difficulties (Section 4.4). Finally, our experiment opened up further research questions, and we provide preliminary guidance on further research paths (Section 4.5).

4.1. Factual learning about objectives

(RQ1) Yes, the proposed gamified interface to inform on objectives facilitated factual learning, as did the nongamified information.

Both the gamified interface and the identical text in the nongamified control treatment facilitated factual learning. Knowledge scores showed that participants significantly increased their knowledge about the wastewater objectives by roughly 25%. This was also supported by the qualitative feedback (Table 3, SI 10a). For instance, the game design students found the gamified interface informative (SI 9a). Thus, the gamified interface is suited to teaching laypeople about objectives relevant to wastewater management. However, contrary to our expectations, the gamified interface led to just as much factual learning as the control.

This lack of difference in learning between the treatments dovetails with fairly equivocal findings regarding the usefulness of gamification to facilitate learning. Some reviews report positive successes (e.g., Kasurinen and Knutas, 2018; Subhash and Cudney, 2018), whereas others have mixed views. For instance, a review paper on social learning with collaborative serious games reports evidence for cognitive and normative learning, but relational learning is not sufficiently investigated to reach conclusive results (den Haan and van der Voort, 2018). One theoretical paper is provocatively titled "why gamification fails in education (…)" (van Roy and Zaman, 2017). Dichev and Dicheva (2017) also report that 64% of 41 reviewed papers are inconclusive, another 10% report negative results, and just 26% report results supporting the usefulness of gamification for learning in higher education. This suggests that gamification can backfire, for instance, if participants are caught in the game and forget the "serious" part, or if participants are overloaded by learning how to play, thereby not being able to process the new information. Given the literature, our results are plausible.

In sum, it is somewhat disappointing that the gamified information on objectives for factual learning is only as good as the control, especially because the results stem from a rigorously designed experiment. However, the gamified interface was as good as the control, suggesting that it was successful in conveying the relevant knowledge, while also being experienced as more entertaining.

In addition, the lack of differences in our study could at least partially be due to our sample composition and testing conditions. Despite recruiting participants broadly (including advertising on a platform for small jobs, social media, by broad word-of-month, and by e-mail to apprentices), 75.5% of the participants had a university degree. Most participants were used to reading and processing long and complex texts. Additionally, the formal experimental situation in cubicles at the university and the monetary incentive may have increased the motivation to perform well, and the sense of obligation to follow the instructions. Further studies should investigate the extent to which individual characteristics and experimental conditions influence the perception of gamification and its outcomes.

The Swiss case study provided preliminary insights on this sampling issue. Only three of 20 respondents in the case study had a higher education; the remaining 17 respondents belonged to our target group, laypeople. The final knowledge score of these 17 respondents did not differ strongly from those with higher education (6.4 and 7.0, respectively) and was about the same as our experimental results (Table 3). The small and unbalanced sample sizes only allow preliminary

observations. Proper experiments with larger public samples and controlling for potential moderating factors, such as education level and learner type (Buckley and Doyle, 2017) would enable better understanding of the effects of gamification (Dichev and Dicheva, 2017; Nacke and Deterding 2017; Landers et al., 2018).

Finally, this study focused on a gamified online interface to inform about objectives of wastewater management. The proposed gamification can be improved to increase the feelings of autonomy, competence, and relatedness (see Sections 4.3 and 4.4). Future research will need to test whether changing aspects of gamification would lead to increased factual learning. Alternatively, other means to enhance factual learning could also be investigated. This endeavor would require reviewing literature for learning in adulthood (Merriam and Baumgartner, 2020), and/or multimedia learning (Niegemann and Heidig, 2012). To mention one example, instructional videos might be a means to inform laypeople about the objectives (e.g., Norman, 2017; Espino et al,. 2021).

4.2. Preference construction and range insensitivity bias

(RQ2) Informing participants on objectives making attribute ranges explicit facilitated preference construction, which, however, was insufficiently range-sensitive.

Our results for the gamified treatment suggest that participants constructed preferences and that the elicited weights reflected their opinion (Section 3.3). Additionally, participants reported considering the best and worst attribute levels. The dedicated test questions partially supported this self-reporting: the participants' perceived range sensitivity was medium on average (about two of four points). Thus, the gamified interface facilitated range-based preference construction to some degree.

However, despite designing our gamified interface to emphasize the attribute ranges, our gamification failed to create stronger range sensitivity than the nongamified information on objectives. One reason for the lack of evidence for higher range sensitivity in the gamified treatment than in the control could be that range sensitivity relates to factual learning, because factual knowledge about the objectives is a prerequisite for considering the ranges. Given that we did not find differences in factual learning (RQ1), the lack of differences in range sensitivity (RQ2) might not be surprising. Alternatively, the way we made the ranges explicit in the dedicated chapter on the worst and best levels (SI 1) could be suboptimal. On hindsight, we are uncertain whether we were able to make the texts about the consequences of the extreme levels of attributes tangible enough to make sense for laypeople. As future research proposition, we suggest explicitly testing different types of textual information as well as other information modes for conveying ranges to laypeople. This could include a small training task at the beginning of the task so that respondents can experience the effects of range insensitivity (Anderson and Clemen, 2013). Moreover, our preliminary search for correlation between knowledge scores and range perception are not conclusive (SI 8b), and require further investigation. For instance, individual RSIs could be calculated, instead of between participants.

The range sensitivity indices (RSI) is the ratio between the expected and elicited weights (von Nitzsch and Weber, 1993). These were rather low in our experiment (at best 0.63), but in line with those found in the literature, which were at best 0.62 (Monat, 2009, table 1). Some of our RSIs

were even negative (Table 5). Negative RSIs do not comply with the range sensitivity principle. Negative RSIs show that an objective receives a higher weight with a narrow attribute range than with a wide one. The weight-change model (Wedell, 1998) could explain this. This model suggests that in a pairwise comparison, when the range of attribute level is extended on Dimension A, the relative weight given to differences along Dimension A decreases, because the weights for both Dimensions A and B are adjusted relative to the larger ratio of difference in the attribute levels (see numerical example in Wedell, 1998). Wedell (1998) tested their weight-change model in a series of three experiments with 159, 98, and 143 participants. However, their results did not support the weight-change model, whereas ours might. If future studies also report negative RSIs, research should test this model in detail.

Low RSIs, indicating insufficient weight adjustments to range manipulation, are common (Monat, 2009). Possibly, we used inadequate linear single-attribute value functions to calculate the RSIs. RSIs depend on the shape of the value function (Weber and Borcherding, 1993) and is congruent with the value-shift model (Wedell, 1998; Verlegh et al., 2002). This model suggests that the participants rescale the value functions to the changed attribute ranges instead of adapting the weights. Accordingly, van Ittersum and Pennings (2012) inferred local weights based on global non-linear value functions. This seems to be a promising avenue, with or without gamification. Future research should continue exploring the reasons and potential remedies to the range insensitivity bias, which is fundamentally important in MCDA, and where we see some urgency for finding effective debiasing methods (Montibeller and von Winterfeldt, 2015).

Concerns about range insensitivity relate to what the weights represent. In particular, the multi-attribute value theory literature discusses global versus local weights. Recent articles suggest relevancy analysis (Marttunen et al., 2019) or the swing weight matrix method (Ewing et al., 2006) to overcome the range insensitivity bias. These approaches view weights as a combination of global importance, which is based on peoples' values, and local importance, which is based on the attribute range in the specific application case. Alternatively, Monat (2009) suggests eliciting first the respondents' own global scale, that is, best and worst attribute levels based on experience or imagination, and second the weights using direct rating. Thereafter, the decision analyst can calculate the weights to fit the local scale. Eliciting global ranges from participants in trade-offs was tested in the nineties (Baron, 1997). However, it may create an anchoring bias (Montibeller and von Winterfeldt, 2015). Further experiments could compare these procedures and the standard trade-off and swing weight elicitation methods.

We can learn from experiences gathered to tackle other biases. For instance, the splitting bias persisted, despite informing about the bias, raising awareness, and educating about the elicitation process and pitfalls (Weber et al., 1988; Pöyhönen and Hämäläinen, 2001; Jacobi and Hobbs, 2007). It might be worth developing debiasing approaches in case the range sensitivity bias also persists, despite providing information that points to avoiding it (Jacobi and Hobbs, 2007). Debiasing approaches could include (1) multiple elicitations from the same person to reveal possible conflicts in answers and ask participants to correct their obviously biased preferences, or (2) mathematically correct the biases, for example, following up on van Ittersum and Pennings (2012). In general, OR researchers could reflect on whether gamification could help overcoming known biases, beyond the already mentioned range insensitivity or anchoring bias. Potentially interesting biases include group think, anchoring, or the status quo bias. Furthermore, it is important to ensure that the gamification does not introduce or strengthen biases. For instance, a wrongly developed storyline

or graphic design could make one objective more salient than others, thereby potentially triggering the availability bias (Montibeller and von Winterfeldt, 2015).

A potential limitation of our study may be the weight elicitation procedure, which is supported by an online interface, but without an experienced decision analyst. We used an improved interface in comparison to previous attempts (Aubert et al., 2020). Its evaluation was positive, but we suggested further improvements (Aubert et al., [accepted]). Online weight elicitation presents both advantages and challenges (Mustajoki et al., 2004; Insua and French, 2010; Aubert et al., 2020). Specifically, online participation — aided with a decision support system, but no experienced analyst — presents the following advantages: space and time constraints no longer apply. Once set by an experienced analyst, less experienced operators can easily carry it out, adapt, or replicate it in different contexts. Being accessible to almost everybody, affected laypeople can be involved, potentially broadening the range of opinions and helping to identify conflicts. Online participation in such surveys enables involving many individuals in decision-making, and satisfies a call for increased participation in praxis (e.g., French and Argyris, 2018; Johnson et al., 2018). However, challenges remain (Marttunen and Hämäläinen, 2008). To date, the digital divide remains a reality. Cognitively demanding and long surveys can be irksome and demotivating. It is unclear whether participants really understand the instructions or the process as a whole (Mustajoki et al., 2004; Introne and Iandoli, 2014), which could increase the risk of people using heuristics and exhibiting biases. Good understanding of the instructions and the process is a prerequisite for high-quality weight elicitation, whether facilitated by an analyst or not. To circumvent the absence of facilitation and the risk of misunderstanding of the process, some researchers favor online interactive group decision and negotiation (Insua and French, 2010). They aim to increase in-depth understanding of the selected representative participants (Bayley and French, 2008; Benyoucef and Verrons, 2008; Vieira et al., 2020).

The MCDA community is exploring how to contribute to e-democracy (Insua and French, 2010). Several paths are possible: one can focus on the democratic process, or on the tools and instruments used in the process (Lavin and Rios Insua, 2010). Researchers can also focus on increasing the understanding of involved participants, or on increasing the number of participants. Online preference elicitation for independent use are tools to increase the number of participants. Few studies are available so far (Mustajoki et al., 2004; Insua and French, 2010; Gregory et al., 2016; Lienert et al., 2016; Aubert et al., 2020). Future research should control specifically for process understanding, biases, and heuristics use, and systematically evaluate the online interfaces. As Phillips (2006, p.188) wrote: "Is an unsound tool better than unaided judgement? One wonders."

4.3. Gamification to improve user experience

(RQ3) The gamified information on objectives was more entertaining.

Comparing items of user experience, the only statistically significant difference between the gamified and nongamified treatments was that the gamified information on objectives was more entertaining than the control (Section 3.4, SI 8c). In all other respects, both treatments received neutral to positive assessments. There are two potential reasons for this lack of a significant effect in the overall experience between treatments. One reason could be at the manipulation level: our items

were reliable, but there was no difference to measure, because the manipulation was too weak. Another reason could be at the measurement level: the manipulation was fine, but the items did not measure reliably enough what we manipulated. Items for the overall experience came from the literature (e.g., Harms et al., 2014). We tend toward the first explanation at the manipulation level, because the differences in the constructs meant to influence the overall experience (the three basic psychological needs) were themselves minor. We discuss this hereafter.

There were no statistically significant differences between the gamified and nongamified treatments for the constructs of autonomy, competence, or items of relatedness (Section 3.4, Table 6). The constructs of autonomy and competence showed acceptable internal reliability and received neutral to positive assessments. However, the internal reliability (Cronbach's alpha) of the relatedness scale suggested issues in our selection of items. We discuss these results in more detail below.

The SDT, and in particular the basic psychological needs theory (BPNT), is one of the most commonly used theories to explain the effect of gamification (Seaborn and Fels, 2015; Nacke and Deterding, 2017; Ryan and Deci, 2017a; Ryan and Rigby, 2019). Several studies have reported positive effects of gamification on the three constructs of the BPNT (Gagné, 2003; Wee and Choong, 2019; Xi and Hamari, 2019). To the best of our knowledge, no scales to measure the constructs of the BPNT for gamification were properly validated yet. Therefore, we adapted those scales that we found in the literature, as these authors did themselves (Sheldon and Filak, 2008; Przybylski et al., 2010; Tamborini et al., 2010; Peng et al., 2012; Ryan and Deci, 2017a; Xi and Hamari, 2019). In practice, one should formulate the question in the same way as the original questions, but ensure that it makes sense in the new context. Given our results (particularly for relatedness, see below), our adjustments might have been too strong, and we recommend reconsidering the original items (Gagné, 2003), or newly partly validated scales for gamification (Högberg et al., 2019).

We found that competence and autonomy were positively correlated with entertainment in both conditions. This indicated that feeling competent and volitional increases the positive perception of the interface. However, perceived competence, that is, how much one feels successful and effective, was statistically not significantly higher in the gamified condition than in the nongamified one, and its relation to entertainment was similar in the gamified and nongamified conditions. We manipulated competence within the gamified interface based on small tasks with clear goals such as matching tasks and progress. Apparently, this was not enough to significantly increase the perceived competence or lead to higher correlation between competence and entertainment.

Autonomy, that is, deciding voluntarily for oneself, also showed no statistically significant differences between the gamified and nongamified conditions. This is unexpected because the gamified interface contained game elements that are usually connected with the experience of autonomy, such as being able to choose the order of events (Ryan and Rigby, 2019). In our interface, the participants could choose the order in which they discovered and read the information about the objectives. Possibly, the inclusion of more game elements connected with autonomy would have produced a different result. Game elements associated with immersion, including narratives and avatars, have a positive effect on perceived autonomy (Xi and Hamari, 2019). Our interface contained a nonplayer character, Mr. Akles, but the lack of a virtual relationship with that character (see next paragraph) suggests that the created environment was insufficiently immersive. Moreover, a nonfixed structure such as a nonlinear storyline might better satisfy the need for autonomy (Wee and Choong, 2019), a feature that was not offered in our gamified interface. Finally, experimental conditions can also restrict the participants' feeling of autonomy (Heimann and Roepstorff, 2018),

which is essential for playfulness. Certainly, sitting in a cubicle in the university lab to conduct a computer experiment is hardly a playful condition. The medium correlation between experienced autonomy and entertainment, however, indicated that it is promising to consider game elements that increase autonomy when developing a game-based interface, as an increased feeling of autonomy leads to a better experience.

Third, we failed to measure a single construct for relatedness, which represents connectedness and exchange with others. Our three items for measuring relatedness (SI 5a) showed unacceptable internal reliability. Further analyses (SI 8e) confirmed that these three items measured different aspects and not one single construct. This could be because relatedness in case of gamified information on objectives for societal decisions might have two aspects: (1) relatedness to the nonplayer characters and (2) relatedness to society as a whole. This dichotomy of relatedness is also discussed in the literature. The players can be related to the nonplayer characters in the game-based interface independently of feeling related to other participants (e.g., in our case, society) (Ryan and Rigby, 2019). Our construct included different subdimensions in one, which turned out to be unsuccessful. A recent new questionnaire to measure the experience of gamification, gamefulquest (Högberg et al., 2019), includes a specific section for social experience, one of the subdimensions of relatedness. However, these authors do not identify the connection with the nonplayer characters as a stand-alone dimension. Investigating gamification with BPNT is an emerging field of research, and we need to develop standards, for instance, for the items and the variables investigated. Our results contribute to this rapidly growing gamification literature. It can serve further research to identify critical issues and support converging toward common standards.

Finally, despite the proposition that relatedness can be experienced through a virtual relation with a nonplayer character (Ryan and Deci, 2017a; Ryan and Rigby, 2019), hardly any research has investigated the importance of nonplayer characters per se (Mekler et al., 2017; Wee and Choong, 2019). Instead, the need for relatedness was satisfied by competition, cooperation in a team, and social networks (Wee and Choong, 2019; Xi and Hamari, 2019). Results relating most to nonplayer characters show that avatars, a meaningful story, and teammates are elements that positively influence social relatedness (Sailer et al., 2017). Most studies identify nonplayer characters as a part of the narrative or meaningful story. Our results showed a lack of virtual relationship and call for specifically investigating the effect of nonplayer characters in gamification. Their design can demand substantial resources, up to using artificial intelligence to adapt the nonplayer character to the player, as already occurs in some video games (e.g., Kopel and Hajas, 2018). This question has already been raised: "we are intrigued by how needs for relatedness may be met by 'computer generated' personalities and artificial intelligence" (Ryan et al., 2006, p. 350). Can participants relate to simple, less costly, nonplayer characters? Should nonplayer characters have specific traits to facilitate virtual relationships? What are these traits? We call for investigating the relatedness to nonplayer characters in addition to social relatedness.

4.4. Limits of the proposed gamification

Our results showed that our gamified information failed to create immersion, and in particular, relatedness to the nonplayer character Mr. Akles. The nonplayer character was a cornerstone of the proposed gamification, and the absence of relatedness to Mr. Akles could explain the overall

lack of effect of our gamification. The nonplayer character and avatar could benefit, for instance, by including possibilities for interactive dialogues that create emotions or at least empathy. These are elements that make a meaningful storyline, which facilitates immersion and in turn the feeling of autonomy (Xi and Hamari, 2019). Additionally, improving voice-over and sound ambiance could contribute to immersion. Qualitative feedback indicated that the voice-over that we used in our interface should be less theatrical and more natural. The whole experience could benefit from background sounds, and each objective could have a specific sound signature in addition to the icon. Finally, more diverse nonplayer characters could be included, each with different opinions. This would highlight the diversity of worldviews while creating options for more interactions and dialogues. Implementing these aspects in a nonlinear structure would also increase the feeling of autonomy (Heimann and Roepstorff, 2018).

Random events are game elements that break the linearity of fixed structures, which we also implemented in our gamified interface as drinking coffee with Mr. Akles (Table 1). These also need improvement, as they should contribute to fulfilling the need for autonomy (Ryan and Rigby, 2019). Currently, the main random events occurred when the participant was inactive to remind them to take action. Thus, participants that remained active did not encounter any random events. If encountered, they had little to no consequences on the fixed storyline, as they just served as reminders. Consequently, game design students judged them as being useless. They suggested relating them to the topic of wastewater management to give additional information and a stronger meaning to the storyline, thus increasing its immersive potential. As for the nonplayer character, we did not find any discussion of the effect of random events per se in the recent gamification literature.

The present study focused on Evaluation (phase VI of the gamification design method we used) (Morschheuser et al., 2017). After this phase, it is common to iterate back to phase IV Design, or even III Ideation, to "solve" the problems revealed in the Evaluation phase. In hindsight, we could have engaged earlier with users via test panels, in shorter and more frequent iteration loops. Co-designing is a common practice when developing serious games for complex decisions, as often done in participatory modelling (e.g., Barreteau et al., 2014; Ditzler et al., 2018). Co-designing usually ensures that the gamification matches the audience, and can reveal "wrong ideas" early in a design process. It is also in line with user-centered design (Hollender et al., 2010). However, it is also highly resource consuming. In some cases, the co-design process is even the aim. Co-design processes help participants to learn and make explicit their tacit knowledge (Barreteau et al., 2014; Aubert et al., 2019).

4.5. Further outlooks for research and gamification

SDT is not the only valid theoretical framework for investigating gamification but it is currently commonly used (Seaborn and Fels, 2015; Nacke and Deterding, 2017; Ryan and Deci, 2017a; Ryan and Rigby, 2019), mostly as the basic psychological needs subtheory. However, Loughrey and Broin (2018) argue that most gamification research currently oversimplifies SDT and ignores other subtheories such as cognitive evaluation, organismic integration, causality orientations, and goal contents (Ryan and Deci, 2017b). Loughrey and Broin (2018) invite follow-up research: (1) to create a taxonomy of game elements and measure whether they are perceived as extrinsically or intrinsically motivating, noting that this will be user and context dependent (causality

orientations); (2) to verify whether individuals perceiving more external regulation are more likely to integrate extrinsic motivational elements such as game elements, as opposed to individuals perceiving more internal regulation (organismic integration); and (3) to differentiate individuals who are intrinsically motivated, as the use of extrinsically motivating elements, even with an intrinsically motivating functional significance, may thwart their existing motivation. This final aspect might have been the reason for the mixed judgment by the voluntarily participating Swiss citizens in our gamified survey in the case study: they were intrinsically motivated people who did not need the extrinsic motivational elements of gamification. In addition, consideration should be given to the interactions between basic needs (Ryan and Deci, 2017b) and the effects of age, level of education, and exposure to games in general on the perception of gamification. This relates to an unexplored question, whether gamification is also suitable for more senior citizens.

Gamification research is growing. This will shed light on whether its promising features are as powerful as expected, but this requires that rigorous evaluation showing negative or null results are also published (thereby overcoming some publication bias towards positive results) (Sridharan and Greenland, 2009). Indeed, potential drawbacks of gamification also need to be investigated. For instance, is gamification diverting attention and cognitive capacity from the main task?

Finally, if the gamification is digital as in our case, the frameworks used in the field of Human-Computer Interaction (HCI) can be relevant (e.g., Carroll, 1997; Hochheiser and Lazar, 2007). HCI among others deals with the usability of programs or digital tools, that is, how users behave and/or accept them. Interestingly, some of this work also refer to some SDT subtheories (Venkatesh, 2000).

5. Conclusion

We developed and thoroughly tested an online interface in a controlled experiment. The aim of the online interface was to inform laypeople about objectives in a real decision about sustainable wastewater management, and to elicit their informed weight preferences for later use in decision-making. This is important because online interfaces for MCDA, which can be used unaided by a decision analyst, could make MCDA processes more participatory. For instance, many affected citizens could take part in the decision process, which so far has been difficult to achieve in decision workshops.

We tested the gamified information on objectives for three aspects. First, factual learning occurred: the participants were better informed about the objectives after completing the survey. However, contrary to our expectations, factual learning in the gamified treatment was not higher than in the nongamified control treatment. This is in line with a growing body of studies on game-based learning that report no evidence of an increase in factual learning due to gamification (Dichev and Dicheva, 2017; van Roy and Zaman, 2017). One reason in our case could be suboptimal gamification design, which we discuss.

Second, range-based preference construction was confirmed. However, our results, in line with the existing literature, showed that the weights assigned to manipulated attribute levels were not sufficiently adjusted. Differences between the gamified and control treatments in range sensitivity were equivocal. Future experimental settings such as within-subject design for the range manipulation may be useful.

Third, the gamified interface to inform about objectives provided a neutral to positive user experience, and was more entertaining than the nongamified control. However, gamification did not increase feelings of competence and autonomy. Our questionnaire, which is fully available for other researchers, requires further development. Our preliminary results on this study are not readily comparable to others in the literature. One reason is that gamification is an emerging field in which no standards exist yet. Our results contribute to ongoing scientific debates about (1) serious games and gamification to enhance learning; (2) preference construction, questioning the range sensitivity principle, and the meaning of weights in MCDA; and (3) evaluating gamification based on the SDT.

Despite the results not always being statistically significant, our study provides interesting insights. We have learnt that our gamified online information on objectives may have suffered from a suboptimal gamification, in particular not sufficiently satisfying the need for autonomy of participants by proposing a too linear story. In addition, the scale to measure the relatedness is insufficient. It should distinguish between the connectedness to the nonplayer character and social connectedness, which is very relevant in public decision-making. Co-designing games in participatory modeling would be another promising approach to support gamified decision-making for complex cases.

We discuss future research lines, which include: (1) continuing efforts to assess the feasibility of online MCDA interfaces that allow for eliciting preferences unaided by a decision analyst; (2) carrying out studies to test different models that could explain range insensitivity, as our results questioned the range sensitivity principle; (3) having a more comprehensive understanding of the SDT in studies focusing on gamification by replicating experiments with different participants; (4) developing gamification-specific validated and reliable basic psychological needs scales; and (5) testing the effectiveness of different interventions. Generally, we targeted a growing societal demand by investigating an innovative way to engage laypeople in complex environmental decision-making while exploring its effect on a well-known bias. We encourage our colleagues, particularly from behavioral operational research, to join us in the exciting expansion of traditional decision support approaches offered by technical online developments.

the case study. Finally, we are grateful for the anonymous reviewers' comments, which helped to greatly improve this paper.

Open access funding provided by ETH-Bereich Forschungsanstalten.

Software availability

Name of software: Edanaga – informing on objectives. Developers: Aubert (concept), Opinion Games (Brugg, Switzerland: https://opiniongames.ch/, retrieved on 15.07.2019) (implementation). Year first available: December 2018. Hardware required: better with a big screen. Software required: Mozilla web-browser. Availability: on demand. Prototype testing: free of charge.

Name of software: Edanaga – online swing weight elicitation. Developers: Aubert and Lienert (concept), Ali Sahandabadi (implementation). Year first available: December 2018. Hardware required: better with a big screen. Software required: Mozilla web-browser. Availability: https://c4science.ch/diffusion/8244/ (Eawag_swing_Weight_Elicitation). Prototype testing: free.

References

Anderson, R.M., Clemen, R., 2013. Toward an improved methodology to construct and reconcile decision analytic preference judgments. *Decision Analysis* 10, 121–134.

Aubert, A.H., Esculier, F., Lienert, J., 2020. Recommendations for online elicitation of swing weights from citizens in environmental decision-making. *Operations Research Perspectives* 7, 100156.

Aubert, A.H., Lienert, J., 2019. Gamified online survey to elicit citizens' preferences and enhance learning for environmental decisions. *Environmental Modelling & Software* 111, 1–12.

Aubert, A.H., Medema, W., Wals, A.E.J., 2019. Towards a framework for designing and assessing game-based approaches for sustainable water governance. *Water MDPI* 11, 869.

Aubert, A.H., Schmid, S., Beutler, P., Lienert, J., Innovative online survey about wastewater management: What young Swiss citizens know and value. Accepted in *Environmental Science and Policy*.

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, 17–28.

Baron, J., 1997. Biases in the quantitative measurement of values for public decisions. *Psychological Bulletin* 122, 72–88.

Barreteau, O., Bousquet, F., Étienne, M., Souchère, V., dAquino, P., Etienne, M., 2014. Companion modelling: a method of adaptive and participatory research. In Étienne, M. (ed.) *Companion Modelling*. Springer, Amsterdam, pp. 13–40.

Bayley, C., French, S., 2008. Designing a participatory process for stakeholder involvement in a societal decision. *Group Decision and Negotiation* 17, 195–210.

Belton, V., Elder, M.D., 1994. Decision support systems: learning from visual interactive modelling. *Decision Support Systems* 12, 355–364.

Belton, V., Stewart, T.S., 2002. *Multiple Criteria Decision Analysis: An Integrated Approach*. Kluwer Academic Publishers, Dordrecht.

Benyoucef, M., Verrons, M.H., 2008. Configurable e-negotiation systems for large scale and transparent decision making. *Group Decision and Negotiation* 17, 211–224.

Beutler, P., Lienert, J., 2020. *Zukünftige Abwasserentsorgung im ländlichen Raum - Fallstudie 2. Abschlussbericht für die Gemeinde*. Dübendorf, Switzerland

Buckley, P., Doyle, E., 2017. Individualising gamification: an investigation of the impact of learning styles and personality traits on the efficacy of gamification using a prediction market. *Computers & Education* 106, 43–55.

Carroll, J.M., 1997. Human–computer interaction: psychology as a science of design. *International Journal of Human-Computer Studies* 46, 501–522.

Cockerill, K., Glynn, P., Chabay, I., Farooque, M., Hämäläinen, R., Miyamoto, B., McKay, P., 2019. Records of engagement and decision making for environmental and socio-ecological challenges. *EURO Journal on Decision Processes* 7, 243–265.

Connolly, T.M., Boyle, E.A., MacArthur, E., Hainey, T., Boyle, J.M., 2012. A systematic literature review of empirical evidence on computer games and serious games. *Computers & Education* 59, 661–686.

Cunliffe, D., 2000. Developing usable web sites – a review and model. *Internet Research* 10, 295–308.

Darejeh, A., Salim, S.S., 2016. Gamification solutions to enhance software user engagement: a systematic review. *International Journal of Human-Computer Interaction* 32, 613–642.

Dede, C., 2009. Immersive interfaces for engagement and learning. *Science* 323, 66.

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, 4529

Deterding, S., 2012. Gamification: designing for motivation. *Interactions* 19, 14–17.

Dichev, C., Dicheva, D., 2017. Gamifying education: what is known, what is believed and what remains uncertain: a critical review. *International Journal of Education Technology in Higher Education* 14, 9.

Ditzler, L. et al., 2018. Affordances of agricultural systems analysis tools: a review and framework to enhance tool design and implementation. *Agricultural Systems* 164, 20–30.

Duke, R.D., 1980. A paradigm for game design. *Simulation & Gaming* 11, 364–377.

Eisenführ, F., Weber, M., Langer, T., 2010. *Rational Decision Making*. Springer, Berlin–Heidelberg.

Ewing, P.L., Tarantino, W., Parnell, G.S., 2006. Use of decision analysis in the army base realignment and closure (BRAC) 2005. *Military Value Analysis Decision Analysis* 3, 33-49.

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, 1149–1160.

Fischer, G.W., 1995. Range sensitivity of attribute weights in multiattribute value models. *Organizational Behavior and Human Decision Processes* 62, 252–266.

Franco, L.A., Hämäläinen, R.P., 2016. Behavioural operational research: returning to the roots of the OR profession. *European Journal of Operational Research* 249, 791–795.

Franco, L.A., Hämäläinen, R.P., Rouwette, E., Leppänen, I., 2021. Taking stock of behavioural OR: a review of behavioural studies with an intervention focus. *European Journal of Operational Research* 293, 401–418.

French, S., Argyris, N., 2018. Decision analysis and political processes decision. *Analysis* 15, 208–222.

Gabrielli, W.F., von Winterfeldt, D., 1978. Are important weights sensitive to the range of alternatives in multiattribute utility measurement? *Research Report* 78.

Gagné, M., 2003. The role of autonomy support and autonomy orientation in prosocial behavior engagement. *Motivation and Emotion* 27, 199–223.

Goldstein, W.M., 1990. Judgements of relative importance in decision making: global vs. local interpretations of subjective weight. *Organizational Behavior and Human Decision Processes* 47, 313–336.

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, 560–565.

Haag, F., Zürcher, S., Lienert, J., 2019. Enhancing the elicitation of diverse decision objectives for public planning. *European Journal of Operational Research* 279, 3, 912–928.

Hamari, J., Koivisto, J., Sarsa, H., 2014. Does Gamification Work? A Literature Review of Empirical Studies On Gamification. In: 47th Hawaii International Conference on System Sciences, Hawaï, USA, 6–9 January 2014, pp. 3025–3034.

Harms, J., Wimmer, C., Kappel, K., Grechenig, T., 2014. Gamification of online surveys: conceptual foundations and a design process based on the MDA framework. Paper presented at the Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational, Helsinki, Finland,

Heimann, K.S., Roepstorff, A., 2018. How playfulness motivates – putative looping effects of autonomy and surprise revealed by micro-phenomenological investigations. *Frontiers in Psychology* 9, 1704.

Hobbs, B.F., Meier, P.M., 1994. Multicriteria methods for resource planning: an experimental comparison. *IEEE Transactions on Power Systems* 9, 1811–1817.

Hochheiser, H., Lazar, J., 2007. HCI and societal issues: a framework for engagement. *International Journal of Human–Computer Interaction* 23, 339–374.

Hoeffler, S., Ariely, D., West, P., Duclos, R., 2013. Preference exploration and learning: the role of intensiveness and extensiveness of experience. *Journal of Consumer Psychology* 23, 330–340.

Högberg, J., Hamari, J., Wästlund, E., 2019. Gameful Experience Questionnaire (GAMEFULQUEST): an instrument for measuring the perceived gamefulness of system use. *User Modeling and User-Adapted Interaction* 29, 619–660.

Hollender, N., Hofmann, C., Deneke, M., Schmitz, B., 2010. Integrating cognitive load theory and concepts of human–computer interaction. *Computers in Human Behavior* 26, 1278–1288.

Huotari, K., Hamari, J., 2017. A definition for gamification: anchoring gamification in the service marketing literature. *Electronic Markets* 27, 21–31.

Insua, D.R., French, S., 2010. *e-Democracy: A Group Decision and Negotiation Perspective*. Springer, Amsterdam.

Introne, J., Iandoli, L., 2014. Improving decision-making performance through argumentation: an argument-based decision support system to compute with evidence. *Decision Support Systems* 64, 79–89.

Jacobi, S.K., Hobbs, B.F., 2007. Quantifying and mitigating the splitting bias and other value tree-induced weighting biases. *Decision Analysis* 4, 194–210.

Johnson, M.P., Midgley, G., Chichirau, G., 2018. Emerging trends and new frontiers in community operational research. *European Journal of Operational Research* 268, 1178–1191.

Kasurinen, J., Knutas, A., 2018. Publication trends in gamification: a systematic mapping study. *Computer Science Review* 27, 33–44

Keeney, R.L., 1992. *Value-Focused Thinking: A Path To Creative Decision-Making*. Harvard University Press, Cambridge, MA.

Keeney, R.L., Raiffa, H., 1976. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. Wiley, New York, NY.

Keusch, F., Zhang, C., 2015. A review of issues in gamified surveys. *Social Science Computer Review* 35, 147–166.

Koivisto, J., Hamari, J., 2014. Demographic differences in perceived benefits from gamification. *Computers in Human Behavior* 35, 179–188.

Koivisto, J., Hamari, J., 2019. The rise of motivational information systems: a review of gamification research. *International Journal of Information Management* 45, 191–210.

Kopel, M., Hajas, T., 2018. Implementing AI for Non-player Characters in 3D Video Games. In Nguyen, N.T., Hoang, D.H., Hong, T.P., Pham, H., Trawinski, B. (eds) *Intelligent Information and Database Systems*. Springer International Publishing Ag, Cham, pp. 610–619.

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, 315–337.

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

Lavin, J.M., Rios Insua, D., 2010. Participatory processes and instruments. In Rios Insua, D., French, S. (eds) *E-Democracy: A Group Decision and Negotiation Perspective*. Springer, Berlin, pp. 31–45.

Lichtenstein, S., Slovic, P., 2006. *The Construction of Preference*. Cambridge University Press New York, NY.

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

Loughrey, K., Broin, D.O., 2018. Are we having fun yet? Misapplying motivation to gamification. In: 2018 IEEE Games, Entertainment, Media Conference (GEM), 15–17 August 2018, pp. 1–9.

Lourenco, R.P., Costa, J.P., 2007. Incorporating citizens' views in local policy decision making processes. *Decision Support Systems* 43, 1499–1511.

Marttunen, M., Belton, V., Lienert, J., 2018. Are objectives hierarchy related biases observed in practice? A meta-analysis of environmental and energy applications of Multi-Criteria Decision. *Analysis European Journal of Operational Research* 265, 178–194.

Marttunen, M., Haag, F., Belton, V., Mustajoki, J., Lienert, J., 2019. Methods to inform the development of concise objectives hierarchies in multi-criteria decision analysis. *European Journal of Operational Research* 277, 604–620.

Marttunen, M., Hämäläinen, R.P., 2008. The decision analysis interview approach in the collaborative management of a large regulated water course. *Environmental Management* 42, 1026–1042.

Meinard, Y. et al., 2021. What is policy analytics? An exploration of 5 years of environmental management applications. *Environmental Management* 67, 886–900.

Mekler, E.D., Brühlmann, F., Tuch, A.N., Opwis, K., 2017. Towards understanding the effects of individual gamification elements on intrinsic motivation and performance. *Computers in Human Behavior* 71, 525–534.

Merriam, S.B., Baumgartner, L.M., 2020. *Learning in Adulthood: A Comprehensive Guide* (4th edn). Jossey-Bass, Hoboken, NJ.

Monat, J.P., 2009. The benefits of global scaling in multi-criteria decision analysis. *Judgment and Decision Making* 4, 492–508.

Montibeller, G., von Winterfeldt, D., 2015. Cognitive and motivational biases in decision and risk analysis. *Risk Analysis* 35, 1230–1251.

Morschheuser, B., Hamari, J., Koivisto, J.,2016. Gamification in crowdsourcing: a review. In 49th Annual Hawaii International Conference on System Sciences (HICSS), Hawaii, USA, January 5-8, 2016.

Morschheuser, B., Werder, K., Hamari, J., Abe, J., 2017. How to gamify? Development of a method for gamification. In 50th Annual Hawaii International Conference on System Sciences (HICSS), Hawaii, USA, January 4-7, 2017.

Morton, A., Fasolo, B., 2009. Behavioural decision theory for multi-criteria decision analysis: a guided tour. *Journal of the Operational Research Society* 60, 268–275.

Mustajoki, J., Hämäläinen, R.P., Marttunen, M., 2004. Participatory multicriteria decision analysis with Web-HIPRE: a case of lake regulation policy. *Environmental Modelling & Software* 19, 537–547.

Nacke, L.E., Deterding, S., 2017. The maturing of gamification research. *Computers in Human Behavior* 71, 450–454.

Niegemann, H.M., Heidig, S., 2012. Multimedia learning. In Seel, N.M. (ed.) *Encyclopedia of the Sciences of Learning*. Springer, New York—Dordrecht—Heidelberg–London, pp. 2372–2375.

Norman, M.K., 2017. Twelve tips for reducing production time and increasing long-term usability of instructional video. *Medical Teacher* 39, 808–812.

Pahl-Wostl, C., Schönborn, A., Willi, N., Muncke, J., Larsen, T.A., 2003. Investigating consumer attitudes towards the new technology of urine separation. *Water Science and Technology* 48, 57–64.

Payne, J.W., Bettman, J.R., Schkade, D.A., 2006. Measuring constructed preferences: towards a building code. In Lichtenstein, S., Slovic, P. (eds) *The Construction Of Preference*. Cambridge University Press, New York, NY, pp. 629–652

Peng, W., Lin, J.-.H., Pfeiffer, K.A., Winn, B., 2012. Need Satisfaction supportive game features as motivational determinants: an experimental study of a self-determination theory guided. *Exergame Media Psychology* 15, 175–196.

Phillips, L.D., 2006. How Raiffa's RAND memo led to a multi-criteria computer program. *Journal of Multi-Criteria Decision Analysis* 14, 185–189.

Plass, J.L., Homer, B.D., Kinzer, C.K., 2015. Foundations of game-based learning. *Educational Psychologist* 50, 258–283.

Pöyhönen, M., Hämäläinen, R.P., 2001. On the convergence of multiattribute weighting methods. *European Journal of Operational Research* 129, 569–585.

Przybylski, A.K., Rigby, C.S., Ryan, R.M., 2010. A motivational model of video game engagement. *Review of General Psychology* 14, 154–166.

R Development Core Team, 2020. *R: A language and environment for statistical computing* R Foundation for Statistical Computing, Vienna, Austria. Available at http://www.R-project.org/ (accessed 10 November 2020).

Ramirez, D., Squire, K., 2014. Gamification and learning. In Walz. S.P., Deterding, S. (eds) *The Gameful World. Approaches, Issues, Applications*. The MIT Press, Cambridge, MA, pp. 629–652.

Ryan, R., 2017. The motivational pull of video game feedback, rules, and social interaction: another self-determination theory approach. *Computers in Human Behavior* 73, 446–450.

Ryan, R., Rigby, C.S., 2019. Motivational foundations of game-based learning. In Plass, J.L., Mayer, R.E., Homer, B.D. (eds) *Handbook of Game-Based Learning*. The MIT Press, Cambridge, MA, pp. 153–176.

Ryan, R.M., Deci, E.L., 2017a. Motivation and need satisfaction in video games and virtual environments. In *Self-Determination Theory: Basic Psychological Needs in Motivation, Development, and Wellness*. Guilford Press, New York, NY, pp. 508–531.

Ryan, R.M., Deci, E.L., 2017b. *Self-Determination Theory: Basic Psychological Needs in Motivation, Development, and Wellness*. The Guilford Press, New York, NY.

Ryan, R.M., Rigby, C.S., Przybylski, A., 2006. The motivational pull of video games: A self-determination theory approach. *Motivation and Emotion* 30, 4, 344–360.

Sailer, M., Hense, J.U., Mayr, S.K., Mandl, H., 2017. How gamification motivates: an experimental study of the effects of specific game design elements on psychological need satisfaction. *Computers in Human Behavior* 69, 371–380.

Santos Espino, J.M., Guerra Artal, C., González Betancor, S.M., 2021. Video lectures: an analysis of their useful life span and sustainable production. *The International Review of Research in Open and Distributed Learning* 22, 99–118.

Seaborn, K., Fels, D.I., 2015. Gamification in theory and action: a survey. *International Journal of Human-Computer Studies* 74, 14–31.

Sheldon, K.M., Filak, V., 2008. Manipulating autonomy, competence, and relatedness support in a game-learning context: new evidence that all three needs matter. *British Journal of Social Psychology* 47, 267–283.

Sridharan, L., Greenland, P., 2009. Editorial policies and publication bias: the importance of negative studies. *Archives of Internal Medicine* 169, 1022–1023.

Subhash, S., Cudney, E.A., 2018. Gamified learning in higher education: a systematic review of the literature. *Computers in Human Behavior* 87, 192–206.

Tamborini, R., Bowman, N.D., Eden, A., Grizzard, M., Organ, A., 2010. Defining media enjoyment as the satisfaction of intrinsic needs. *Journal of Communication* 60, 758–777.

van Ittersum, K., Pennings, J.M.E., 2012. Attribute-value functions as global interpretations of attribute importance. *Organizational Behavior and Human Decision Processes* 119, 89102.

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: Ma M.O. (ed.) *Serious Games and Edutainment Applications*. Springer, Cham, pp. 485–509.

Venkatesh, V., 2000. Determinants of perceived ease of use: Integrating control, intrinsic motivation, and emotion into the technology acceptance model. *Information Systems Research* 11, 342–365.

Verlegh, P.W.J., Schifferstein, H.N.J., Wittink, D.R., 2002. Range and number-of-levels effects in derived and stated measures of attribute importance. *Marketing Letters* 13, 41–52.

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, 102047.

von Nitzsch, R., Weber, M., 1993. The effect of attribute ranges on weights in multiattribute utility measurements. *Management Science* 39, 937–943.

von Winterfeldt, D., Edwards, W., 1986. *Decision Analysis and Behavioral Research*. Cambridge University Press, Cambridge.

Weber, M., Borcherding, K., 1993. Behavioral influences on weight judgments in multiattribute decision making. *European Journal of Operational Research* 67, 1–12.

Weber, M., Eisenführ, F., von Winterfeldt, D., 1988. The effects of splitting attributes on weights in multiattribute utility measurement. *Management Science* 34, 431–445.

Wedell, D.H., 1998. Testing models of trade-off contrast in pairwise choice. *Journal of Experimental Psychology: Human Perception and Performance* 24, 49–65.

Wee, S.-.C., Choong, W.-.W., 2019. Gamification: predicting the effectiveness of variety game design elements to intrinsically motivate users' energy conservation behaviour. *Journal of Environmental Management* 233, 97–106.

White, L., Bourne, H., 2007. Voices and values: Linking values with participation in ORMS in public policy making. *Omega-International Journal of Management Science* 35, 588–603.

Xi, N., Hamari, J., 2019. Does gamification satisfy needs? A study on the relationship between gamification features and intrinsic need satisfaction. *International Journal of Information Management* 46, 210–221.