European Journal of Operational Research 308 (2023) 1268–1285

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Decision Support

Preference change in stakeholder group-decision processes in the public sector: Extent, causes and implications ☆,☆☆

M. Kuller*, P. Beutler, J. Lienert

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

A R T I C L E I N F O

Article history:
Received 14 February 2022
Accepted 2 December 2022
Available online 7 December 2022

Keywords:
Behavioural operational research (OR)
Preference change
Multi-criteria decision analysis
Group decision-making
Public policy

A B S T R A C T

Public decisions are typically related to large investments leaving long legacies. We should therefore strive for wide societal agreement regarding such decisions, which meet the diversity of preferences between stakeholders and over time. But if, how and why do stakeholder preferences change over time? In decision analysis, these questions received little attention. We explored them using three real-world public decision processes, based on Multi-Criteria Decision Analysis (MCDA). We used repeatedly elicited ranking of objectives over time. These were obtained during three to five moderated workshops we organised several months apart (total $N = 200$ questionnaires, and 100 stakeholders). We analysed individual and aggregated (group) preferences, their changes and potential drivers including demographic and experience variables. We also analysed the effect of preference evolution on the performance of decision-alternatives with MCDA over time. We found that stakeholder preferences often changed over time, both on an individual and group level. These changes did not systematically diminish over time, but some convergence of preferences was observed for stakeholders who repeatedly participated in workshops. High-ranking objectives were relatively stable and similar between stakeholders. While preference changes could not be explained by demographics and personal experiences, repeated interaction with the decision problem might play a role. Neither the observed disagreement between stakeholders, nor the preference changes over time affected the best and worst performing alternatives in our decision problems. Thus, despite changing stakeholder preferences over time, public decision-makers can contrive robust solutions to complex public decision problems in the present.

1. Introduction

Decisions taken in the public sector typically are complex involving many different stakes and stakeholders with diverging preferences. Furthermore, they often leave a long-term legacy. For example, large-scale public infrastructures such as train lines and sewer systems that were decided upon by governments over a century ago still shape our society today. Diligent group decision-making processes are thus paramount to foster effective, robust, fair and consensual decisions from various stakeholders that benefit society in its broadest definition in the long term. Theories and methodologies from Operational Research (OR) including Multi-Criteria Decision Analysis (MCDA) and specifically Multi Attribute Value Theory (MAVT; Keeney & Raiffa, 1976) have been applied in the public sector to support decision-making processes. Such methodologies are designed to reflect underlying preferences of all involved and affected stakeholders as well as possible. The central role of stakeholder preferences in many OR methodologies make them rely on the premise that human preferences regarding the often complex decision parameters exist or can at least be constructed. Additionally, decisions with long legacies start from the assumption that preferences remain relatively unchanged over time. This is especially relevant for those decisions that cannot be revoked without considerable social costs and losses (Gregory et al., 2012), such as the decisions taken in the public sector.

In this study, we used workshop interventions in three exemplary case studies for public sector decisions: they are complex, include uncertainties, stakeholders involved and/or affected by the decisions and they cover long time ranges. One case was carried out in West Africa with the aim of co-developing a flood forecast and early warning system together with West African stakeholders and organisations (EU Horizon 2020 project FANFAR: Andersson et al., 2020; Fanfar, 2021, see Section 2.3.1; Lienert et al., 2022). Flooding is a rapidly growing concern in West Africa, and improved flood management is urgently needed, not least because of cli-

☆ Declaration of interest
☆☆ None

* Corresponding author.
E-mail address: martijnkuller@gmail.com (M. Kuller).

https://doi.org/10.1016/j.ejor.2022.12.001

mate change impacts (Nka et al., 2015). The other two cases were carried out in smaller communities in Switzerland. They concern the transition from the current grid-based centralised wastewater management system to non-grid decentralised options, where wastewater treatment is shifted to individual houses. There are many reasons for this paradigm change, including various advantages at local scale (see Section 2.3.2). Internationally, such a transition from the grid-based central wastewater system to non-grid decentralised systems is receiving growing recognition because it allows addressing pressing global challenges such as rapid urbanization, eutrophication and climate change effects (e.g., Hoffmann et al., 2020; Larsen, et al., 2016).

1.1. Preference construction

A large body of evidence from psychologists convincingly shows that preferences are influenced by various factors (Lichtenstein & Slovic, 2006; Payne et al., 1992). Some of the preference construction processes seem random (e.g., Ariely et al., 2003). Preference construction is likely influenced by socio-economic characteristics and by past experience. For instance, experienced decision makers may retrieve existing preferences if they recurrently face relatively stable situations; they may then be less susceptible to framing effects and more likely to use previously obtained information (reviewed in e.g., Bettman et al., 2008; Warren et al., 2011). Behavioural psychologists explain preference construction with various theories that reflect the many ways in which people process information (see Lichtenstein & Slovic, 2006). Many responses are immediate, and a large body of empirical and experimental research indicates that people use heuristics and simple decision rules or decision strategies (Bettman et al., 1998; Gigerenzer et al., 2011). Social psychology aims to define and understand more fundamental “attitudes” that may be underlying expressions of preferences. Attitude can be defined as “a psychological tendency that is expressed by evaluating a particular entity with some degree of favour or disfavour” (Eagly & Chaiken, 1993, p. 1). This tendency can be understood to be shaped by processes of experience, nature and nurture leaving a mental residue. The evaluations constituting attitude are expressed through cognition, affect and behaviour (Eagly & Chaiken, 2007). We thus understand preferences to be generated by attitudes when an evaluation of the attitude object is expressed implicitly or explicitly.

Human decision-making behaviour can be guided by emotions, affect or memory; and this is an active research field in behavioural decision analysis (e.g., Angie et al., 2011; Slovic et al., 2007). Overall, any preference information must be interpreted in terms of its relation to demographics, the specific decision context, the point in time of the elicitation, the applied elicitation setting (e.g., group vs. individual) and method, the current level of decision-makers' subject-related knowledge and any external influences up to the point of measurement, including experiences (Zheng & Lienert, 2018).

Indeed, most prescriptive decision analysts claim that preferences are constructed during elicitation processes (e.g., Anderson & Clemen, 2013; Belton & Stewart, 2002; Eisenführ et al., 2010; Gregory et al., 2012; Gregory et al., 1993), based on own experience and the insights from behavioural decision theory (e.g., Lichtenstein & Slovic, 2006). Recently, there has been an increased interest in behavioural issues in OR (Franco et al., 2021; Hämäläinen et al., 2013). Behavioural OR covers a broad spectrum of issues, which can be divided into two streams (Franco & Hämäläinen, 2016): (1) concerning the use of OR methods to model human behaviour in complex settings and (2) investigating behavioural aspects in OR to support decision-making and problem solving.

Some specialised fields within OR have been concerned with behavioural issues for decades, namely decision analysis (Von Winterfeldt & Edwards, 1993) and MCDA (French et al., 1998; Korhonen & Wallenius, 1996). However, most of the seminal studies were carried out some 25 to 35 years ago; with an astonishing gap until only relatively recently (e.g., Franco & Hämäläinen, 2016; Franco et al., 2021; Montibeller & Von Winterfeldt, 2015; Morton & Fasolo, 2009). Although it has been shown that manifold variables influence preference construction, we are still far from understanding what is most important to guide preference elicitation processes in real-world, complex public policy decisions. Indeed, a recent review postulates that more real-world studies concerning the effects of individual differences and the impact of e.g., demographic background is needed (Franco et al., 2021). Many of the psychological, consumer and MCDA studies concerning preference construction were carried out in experimental settings, with students, using constructed, relatively trivial decisions, or focusing on smaller consumer choice problems. Largely lacking is scientifically rigorous, empirical work in different real-world decision-making contexts. The evidence that human preferences are constructed and may be susceptible to arbitrary influences is currently not appropriately accounted for in the MCDA methods.

1.2. Preference change

Stated preference studies are static in nature in the sense that they capture only a snapshot of respondents' preferences. Therefore, by far the most commonly used approach to assess preference change over time is test-retest, in which either the same or different samples of respondents are asked exactly the same valuation question at (usually two) different points in time. The majority of studies on preference change come from economics, where test-retest studies are carried out in the context of contingent valuation (CV) and to a lesser extent discrete choice (DC) experiments. Most CV and some DC studies show that preference change is negligible over time and preferences are thus robust (e.g., Berrens et al., 2000; Brouwer, 2006, 2012; Brouwer et al., 2009, 2017; Cameron, 1997; Carson et al., 1997; Fetene et al., 2014; Loomis, 1989; McConnell et al., 1998; Whitehead & Hoban, 1999). However, empirical evidence for discrete choice experiments is more limited and ambiguous, showing that individual choice attributes are insensitive over time, but aggregated values across all attributes tend to differ more substantially (e.g., Bliem et al., 2012; Liebe et al., 2012; Mørkbak & Olsen, 2015; Schaafsma et al., 2014).

From a social psychological perspective, the instability of stated preferences, or the expressed evaluation towards an attitude object, has a different origin from the (in)stability of the attitudes themselves (Eagly & Chaiken, 2007). Preferences are emerging incomplete, imperfect and situationally dependant expressions of attitude, and are therefore more volatile than the attitude. Furthermore, attitudes “tested” (i.e. elicited) in artificial research settings can be more trivial, thus lacking solid mental residue and resulting in increased volatility (Eagly & Chaiken, 2007). Attitudes themselves change either through a “central route” requiring cognitively demanding evaluation of information or “peripheral route”, relying on low-effort short-cuts (Petty et al., 1997). Factors at play are external, such as credibility of a message or piece of information, and personal characteristics such as prior knowledge, socio-economic minority status and mood.

The MCDA literature has largely ignored preference change; for instance it is not mentioned in important earlier reviews (e.g., Hämäläinen et al., 2013). In a recent behavioural OR review, Franco et al. (2021) discuss process and variance studies. These may focus on the development of actor's interactions with the material of an OR intervention over time (Ormerod, 2014), or inversely on the impact of an actor's behaviour on an OR intervention over time. However, *preference change* over time was not specifically mentioned in this review. Empirical observations of preference change in real-world contexts are necessary, as individual preferences from lab-based (consumer) experiments might not translate directly to real decision-making processes (see e.g., Franco et al., 2021). Some recent studies started looking at this; Lienert et al (2016) investigated preference changes of three stakeholder groups, including a population survey, using different elicitation methods for MAVT in the context of public decision-making for Swiss water infrastructure planning. Hayashi et al. (2016) conducted a workshop experiment with municipal officials in Japan on renewable energy using MAVT. The results indicate that increased and repeated exposure to information and the decision support method changed the participants' preferences at least in some cases. This needs verification, since sample sizes were mostly small and limited to specific stakeholder groups and decision contexts.

In line with our knowledge on the construction of preferences, we may expect change of stated preferences to be driven by a range of variables including: a) learning about the system, the decision problem, and personal preferences due to repeated exposure; b) elicitation methods used and provided information; c) external factors and events before and between measuring moments; d) personal characteristics including values, life path, demographics, cognitive style/psychological attributes (including risk attitudes) and individual experience; and e) unknown or latent variables. As emphasised by Gregory et al. (2012), preference change seems especially important for choices that affect long time ranges. This applies to high investments in infrastructures and fundamental system changes that cannot easily be reversed. Ideally, stakeholder preference change is reduced as much as possible, even when faced with moderate changes in context, as a result of a balanced, sufficiently diversified information basis and awareness of unavoidable trade-offs. Therefore, we need to improve our understanding on preferences construction over longer time ranges and the variables influencing preference change.

1.3. Research aim: empirically assessing preference construction and change

As preferences are constructed, they are uncertain and might be subject to change. However, they are valuable and are central to MCDA processes for their role in encouraging learning about stakeholders' own perspectives, as well as for shaping decisions. This study empirically examines behavioural issues in Operational Research over time using repeated interventions, regarding stakeholder preferences towards objectives and alternatives in three distinctively different, real-world decision case studies: co-design of a flood forecast and early warning system in West Africa and wastewater system transitions in a small and middle-sized Swiss village.

Complex public policy problems, which involve many stakeholders, ultimately call for consensus to be reached on a decision. MCDA can support reaching a compromise solution that captures the values and preferences of a wide range of stakeholders, including minorities. However, as stressed before, the robustness of such value-based decision-making processes is essential. We aim to better understand the factors involved in the construction and evolution of preferences and the extent of preference change amongst stakeholders. Three groups of factors investigated are static and include demographic variables, factors related to personal experiences with flood and possible other, unknown (latent) factors that we did not directly measure. We assess whether these factors are related to the nature of stakeholder preferences, as well as their impact on the likelihood that individual preferences change over time. We strive to uncover whether the decision-making process over time – with repeated interventions – help to reduce the divergence in opinions between stakeholders and to increase confidence in our decisions. The objectives of this research are to explore the construction and change of stakeholder preferences over time. We use a simple tool to capture the importance of decision objectives as our preference parameters (Section 2.2). Using regular interventions, we explore preferences on an (1) individual level and (2) on a group level, and (3) analyse the effect of such preference construction- and change on the potential results of an MCDA process (i.e., ranking of alternatives). Connected to these objectives, we pose the following research questions (RQ):

1. Individual stakeholder preferences
   a. What is the relationship between dependant variable individual stakeholder preferences, and independent demographic variables, personal experience and unknown (latent) variables?
   b. Do we observe significant patterns of change in preference rankings of objectives for individual stakeholders over time with interventions?
   c. What is the relationship between the dependant variable "likelihood of a change in individual stakeholder preferences" (if any) and the static independent demographic variables, personal experiences, and unknown (latent) variables?
2. Group stakeholder preferences
   a. Do we observe convergence between stakeholder preferences (i.e., more agreement between stakeholders)? Does convergence differ between respondents who attended all workshops (hereafter called measuring moments, MM), and those who did not?
   b. How do aggregated (group) preferences evolve over time? Is there a difference between the subset of respondents who attended all MM of their case study, compared to the entire set of respondents?
3. What is the impact of changing preferences on the outcome of an MCDA (i.e., ranking and overall value of alternatives) at an individual and group level in three real-world case studies?
   Our research investigates these research questions for each of our three case studies. Additionally, we compare our case studies based on the following research question:
4. How do the answers to the above questions compare between case studies? Do we find evidence for universality of findings between different decision contexts?

Guide for the time-constrained reader: Fig. 1 summarises data gathering, Table 1 presents the research questions, hypotheses and analysis methods. The results are summarised in Table 3. The remainder of this paper is structured as follows: Section 2 presents the data collection approach across the case studies, and the methods of data analysis for each research question. A summary is provided in a table at the end of the section. Results are presented in Section 3, individually per research question, and summarised and compared between case studies at the end of the section. Section 4 discusses the results and methods per research question, presents a general discussion and ends with some suggestions for further research. Section 5 concludes.

2. Data and methods

We attempted to answer our research questions using three very different case studies of real-world decision-making contexts (Section 2.2): (1) co-development of a flood forecast and early warning system (FEWS) in West-Africa (one case study; CS1_Africa) and (2) transitions to future decentralised wastewater systems for rural municipalities in Switzerland (two case studies; CS2_Small and CS3_Larger). To enhance comparability, we applied similar approaches of data collection (Section 2.2) and analysis (Section 2.3) for all three case studies (Fig. 1).

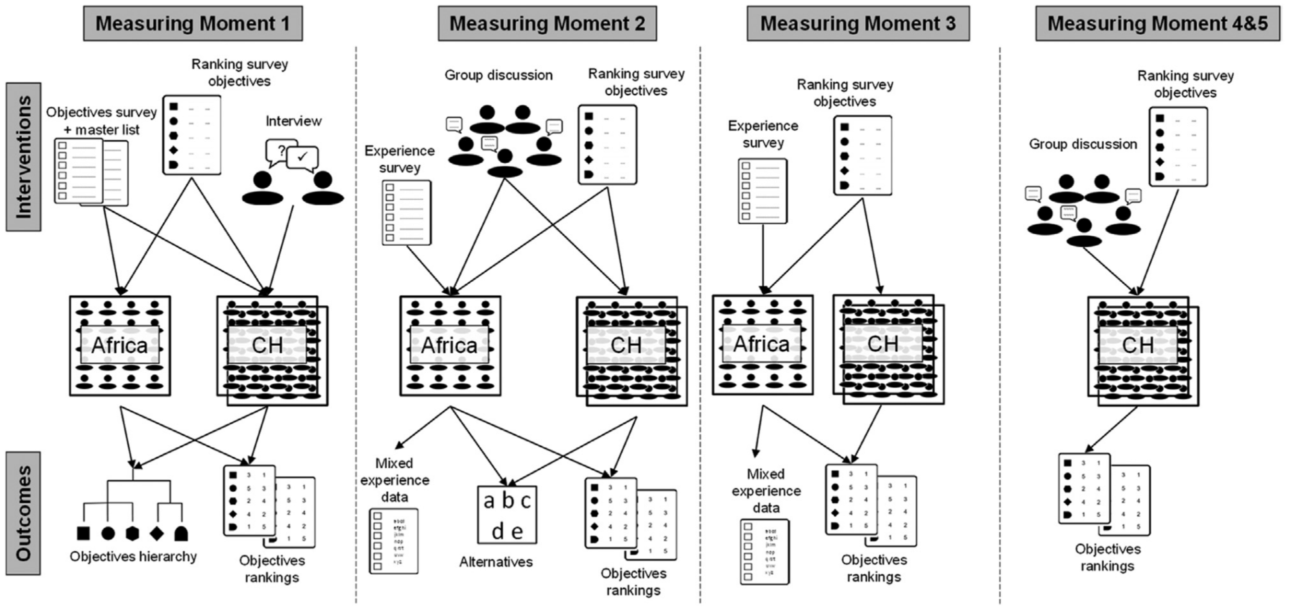

Fig. 1. graphical overview of our data collection approach, connecting all interventions per MM to the three case studies, and in turn relating those to the specific datasets gathered that we used for this article. "Africa" represents CS1_Africa, "CH" represents both CS2_CH_Small and CS3_CH_Larger. All case studies followed a similar procedure, but there are a few differences, particularly between CS1_Africa versus CS2_CH_Small and CS2_CH_Larger. Most notably, the number of MM varied, and the demographics survey was only conducted for CS1_Africa. Details on respondent numbers, split by MM and demographic variables are presented in Fig. SI-1.

Table 1
Summary of the research approach per research question. MM: measuring moments (i.e., workshops).

#	Research question	Working hypotheses	Methods	Data
1a	Relationship between individual preferences and various known and unknown explanatory variables	No relationship with known or unknown explanatory variables	Ordinal regression and cluster analysis	Demographics survey, rankings of objectives
1b	Individual preferences over time	Reduced preference change over time	Kendall's rank correlation (τ), student t-test, visual analysis	Rankings of objectives
1c	What influences the likelihood of individual preference change over time?	No consistent relationship between preference change and demographic or experience variables	Linear regression with β-distribution of Kendall's τ (change) with demographic variables	Demographics survey, rankings of objectives
2a	Do the individual preference sets converge?	Convergence because of group processes and biases; higher convergence for those who visited all MM	Boxplots to visualise spread of rankings and Kendall's coefficient of concordance (W)	Rankings of objectives
2b	How do aggregated preferences evolve over time?	Reduced preference change over time, caused by diminishing individual preference change and by group processes	Friedman test, Friedman test post-hoc analysis, Kendall's τ (between aggregated rankings)	Rankings of objectives
3	What is the impact of changing preferences on the outcome of an MCDA at an individual and group level in three case studies?	The ranking of alternatives stabilises over time, and top ranked alternatives are more stable over time than lower ranked ones	Transformation of ranks into weights using the rank sum (RS) method; MCDA using the ValueDecisions app	Rankings of objectives
4	Is there evidence that the findings from RQ1–3 apply across case studies, i.e., that they may be universal?	Preference formation in the human mind is similar across contexts, thus we expect similar patterns	For each RQ, systematically compare the results between case studies	Results from RQ1–3

2.1. Multi-criteria decision analysis (MCDA)

For all three case studies, we followed standard MCDA procedures, and in particular MAVT (Eisenführ et al., 2010; Keeney & Raiffa, 1976). We started with a problem structuring phase, in which we conducted stakeholder analysis to identify the respondents to include in our co-development process (following the approach by Lienert et al. 2013). We then brainstormed with these respondents during the first workshop to identify what is really important to them, reflected in the agreement on the objectives. Part of the MCDA process is the development of an objectives hierarchy, containing top-level and lower-level (specific) objectives. The extent to which an alternative meets each objective can be measured through attributes (e.g., the degree to which an alternative meets the objective "low cost" can be measured through the attribute "costs" in $/year). We developed the objectives hierarchy during several interactions with the respondents, which we consolidated through a stakeholder group discussion. We then identified possible FEWS configurations (alternatives), which might fulfil these objectives. To predict how well each alternative fulfilled each objective, we interviewed experts. We included different preference information from stakeholders. We elicited marginal value functions and used Swing to elicit preferences regarding the importance of the objectives, i.e., weights (Eisenführ et al., 2010). Additionally, we checked with simple questions whether stakeholders agreed with the additive aggregation model, and because they did not, we used non-additive aggregation to build the overall value function (see e.g., Haag et al., 2019a). The MCDA methods are described in a separate paper (Lienert et al., 2022). However, for the purpose of this study, we used repeated preference elicitation using fast and frugal elicitation methods, focusing only on the ranking of objectives as main preference parameter.

2.2. Fast elicitation of preferences concerning the importance of objectives

Additional to robust eliciting objectives weights using standard Swing in a group setting, we elicited preferences of individual respondents regarding the importance of objectives during each workshop to monitor the evolution of preferences over time (Fig. 1: Measuring Moment). As preference parameters we focused exclusively on the ranking of decision objectives for practical reasons, and because sensitivity analyses had indicated that shapes of marginal value functions were less important (Beutler et al., 2021; Lienert et al., 2022). We used a "fast preference survey" for preference elicitation during all MM to monitor the evolution of preferences over time, using direct ranking and rating of objectives. The core data for the analysis in this paper are the rankings of objectives. To elicit these, we used a pen-and-paper ranking/rating survey, where respondents were asked to rank each objective from 1 (most important) to n (least important) as well as provide a rating. For the latter we used Likert scales (different in each case study). While objectives ranking is a standard procedure of several recognised MCDA techniques (e.g., Swing), direct rating of objectives has been subject to criticism (Eisenführ et al., 2010). In the light of this criticism, and with the finding that the information from rankings is not significantly different from that obtained by rating (Moors et al., 2016), we opted to focus solely on the ranking of objectives. The ranking and rating surveys used for the interventions are presented in SI-1.5.2, SI-1.6 and SI-1.7 (all surveys follow a similar setup, but include different objectives depending on the case study).

2.3. Data collection

We facilitated all workshop sessions (MM) and materials in the local languages (English, French or German), to allow respondents using a familiar language. To enable meaningful comparison between case studies, we designed data collection methods similar but not identical between our case studies (Fig. 1). Data collection took place at several MM in all case studies, ranging from three to five MM. We refer to MM as a collection of interventions with respondents usually taking place as part of a workshop, including face-to-face interviews, team meetings and surveys. At each MM instance, we collected one or several of the following data (Fig. 1): (i) objectives to be considered in the decision-making, (ii) individual respondent's ranking and rating of objectives, (iii) individual respondent's ranking of alternatives, (iv) statements given by respondents during group discussions, (v) respondent characteristics and experiences. Below, we describe the data collection in detail per case study. An overview of the number of respondents for each case study and MM, including demographic and experience variables (when available) are provided in the SI-1.1. Over all three cases, 100 individual stakeholders participated in this study, and we collected and analysed a total of 200 questionnaires.

2.3.1. CS1_Africa

As part of the EU Horizon 2020 project FANFAR (Andersson et al., 2020; Fanfar, 2021), we co-developed a fluvial flood forecast and early warning system (FEWS) with a consortium of European and African organisations (responsible for the technical development and project management) and the hydrology and emergency management agencies (prospective end-users, referred to as respondents for the remainder of this paper) of 17 countries in West Africa. The hydrological forecast system consists of several components. For users, the most important one is the Interactive Visualisation Portal: the main online interface to communicate flood risk forecasts (https://fanfar.eu/ivp/). The system intends to help local authorities and civil society in West African countries to better prepare for incumbent flood events. We designed a co-development approach with the aim to maximise the usefulness of the forecast and warning system to its users.

We conducted a detailed MCDA process over the course of the first and second workshop in order to evaluate respondents' needs and values and prioritise aspects of FEWS development (Lienert et al., 2022). During this process, we identified ten fundamental objectives, which we organised in an objectives hierarchy (Fig. SI-1). Furthermore, we developed 11 FEWS alternatives, using a strategy development table (for details, refer to Lienert et al. (2022). The MCDA included processes of structured and systematic collection of preferences on objectives (amongst other interactions: SI-1.5.2). Additionally, users were able to interact with the FEWS system in development in and between workshops, and we systematically collected experiences with flood events and feedback on technical components of the FEWS (SI-1.5.3). For this study, we only used MCDA elements and feedback on experiences. The consortium organised co-development through three week-long workshops in West Africa (17–20 Sept. 2018 in Niamey, Niger, 9–12 April 2019 in Accra, Ghana, and 10–14 Feb. 2020 in Abuja, Nigeria), complemented by continuous formalised interactions between these workshops (e.g., through online surveys). These workshops represent the MM of this case study. Unfortunately, a fourth workshop (MM4) was cancelled due to the Covid pandemic. During each of the three MM, we elicited direct rankings and ratings using the fast preference survey (SI-1.5.2). During MM1, objectives were developed, and their rankings and ratings elicited simultaneously in three distinct groups: 1) pen and paper survey (similar to elicitation during MM2 and MM3), 2) online survey and 3) group discussion. To select the objectives in MM1, we used a master list of objectives, which could be complemented with additional objectives by respondents (Haag et al., 2019c). This causes the data to show identical, lowest rankings for the objectives that respondents did not identify. While respondents agreed on nine objectives in MM1, during a plenary discussion, we added a 10th objective (*long-term financing secured, 41_sust_financing*) in the consolidation phase before MM2. These 10 objectives (Fig. SI-1) were agreed upon by all participants of MM2 and were used in MM2 and MM3. For the purpose of this analysis, we consequently assigned a ranking of 10 for the additional objective in the data for MM1.

Additionally, two pen-and-paper surveys were taken from the respondents to monitor their experiences with using the FEWS: one during the second (MM2) and one during the third workshop (MM3). Besides questions around FEWS user experiences (not used for this study), the survey included questions about respondents' personal characteristics: language, gender, age group, field of work and years of experience with floods and flood management. We therefore refer to this survey as the demographics survey. The survey consisted of a mix of question types: multiple choice, open and Likert scale (SI-1.5.3).

2.3.2. CS2_CH_small and CS3_CH_larger

In a transdisciplinary research project between Eawag, the Environmental office of the canton of Solothurn and two case study communities, we conducted a complete decision support process based on MCDA. One community consists of fewer than 200 inhabitants (CH2_CH_Small), the other counts 2500 people

(CS3_CH_Larger). Both communities operate a centralised wastewater system based on a sewer network that discharges wastewater from households to a local wastewater treatment plant. High investments are needed for rehabilitation measures of these ageing systems. The high costs raised the question whether it is wise to rehabilitate the existing system, or whether transitioning to an alternative wastewater system would be possible. Non-grid decentralised systems and/or hybrid wastewater systems (grid network enhanced with non-grid elements) could be considered as viable alternatives for various reasons, including increased flexibility and sustainability (Hoffmann et al., 2020; Larsen et al., 2016). Decentralised alternatives have good implementation potential, particularly in low density settlements (Eggimann et al., 2015).

We aimed to support these exemplary communities in rural areas to identify and decide on their future wastewater system matching local conditions and community needs. We facilitated all interventions and provided materials in German. Our approach and methods were applied in both municipalities and the case studies specific results as well as conclusions and recommendations for action are described in detail elsewhere in German publications (Beutler et al., 2021).

We initiated the MCDA process in CS2_CH_Small and repeated it similarly for CS3_CH_Larger. The only difference was that we omitted face-to-face interviews at the start of CS3_CH_Larger. In CS2_CH_Small, we conducted interviews for initial screening of the respondents and to elicit candidate objectives and alternatives. We organised a kick-off event in both cases to narrow down the decision framework and list potential objectives and alternatives. Subsequently, individual respondents completed online surveys based on a master list of objectives (Haag et al., 2019c) to develop a draft objectives hierarchy, which was discussed and agreed upon at the first workshop (MM1; Figs. SI-2 and SI-3). In both of the Swiss case studies, we identified 14 fundamental objectives, of which 11 were identical. Adapted to local conditions, we used sets of nine and 11 conceptual alternatives (CS2_CH_Small and CS3_CH_Larger, respectively; Tables SI-2 and SI-3). At the end of MM1, respondents filled out the fast preference surveys (SI-1.6 and SI-1.7) for the first time, after we had conducted preference elicitation in groups.

Between MM1 and MM2, we only provided further technical details on alternatives (fact sheets) to respondents in CS3_CH_Larger. We learned from CS2_CH_Small that this could support respondents in better understanding the alternatives' implications. Respondents filled out the fast preference surveys at the beginning (MM2) and at the end (MM3) of a second workshop. In this workshop, we presented and intensively discussed MCDA results in groups and in the plenum. Additionally, in CS2_CH_Small, we discussed potential management and operation strategies for the wastewater system alternatives. In CS3_CH_Larger, we discussed concrete transition strategies from central to non-grid decentral wastewater systems for the community. Following MM3, respondents received a written report with detailed and summarised results and decision recommendations for their community. In CS2_CH_Small, we conducted an additional project team meeting without community representatives (MM4) at which respondents filled in the fast preference surveys again. At a last instance (CS2_CH_Small: MM5; CS3_CH_Larger: MM4), four representatives of each community participated in a final, joint workshop. We introduced the two case study project teams to each other and presented their respective case study results, allowing them to learn from each other. At the end of MM5/MM4, respondents conducted the fast preference surveys one last time.

2.4. Data analysis

We designed data analysis after defining detailed working hypotheses (summary provided in Table 1) for each RQ, which are presented in Appendix A. Data of both case studies were analysed similarly, unless specified otherwise. Our core data are the rankings of objectives. While rankings are found to contain similar information to ratings, in-depth experimental comparison showed that they are superior in several aspects including: avoidance of the agreement response style (respondents tend to agree with item regardless of its content) and non-differentiation bias (respondents tend to provide similar ratings to items) as well as providing data with higher information content (e.g., discriminatory power), and higher validity (Moors et al., 2016). Unless specified otherwise, all data manipulation, statistical analyses and graph development were performed using R scripts (R Core Team, 2020), which can be found in Section SI-1.8. The remainder of this section describes the data analysis for each RQ. It ends with an overview of all RQ with associated hypotheses, methods and data used in Table 1.

RQ1a Relationship between individual preferences and explanatory variables. Firstly, we analysed the relationships between the measured independent variables from the demographics survey (CS1_Africa only) and respondents' preferences from the rankings of objectives of all three MM using ordinal regression (McCullagh, 1980). Ordinal regression is an effective method to assess significance, strength and direction of the predictive value of a set of independent variables and the dependant, ordinal variable. We performed this analysis only for CS1_Africa, as we did not measure independent variables for the two Swiss cases. We developed ten separate ordinal regression models, i.e., one for the ranking of each of ten objectives as the dependant variable. The independent variables included: *language* (language spoken by respondent; either English or French), *gender, age group* (low: <35, medium: 35 – 55, high: >55), *field of work* (hydrological services or emergency response services) and *experience* (with flood management in years; low: 1–3, medium: 4–7, high: >8).

Secondly, we explored the presence of latent variables predicting rankings of objectives through cluster analysis (all three case studies). The insertion sorting rank model (ISR) enables the analysis of full ranking data, i.e., entire rank order of a respondent rather than the rank of individual objectives (Biernacki & Jacques, 2013). We applied a model-based clustering algorithm based on the ISR using the Rankcluster package in R (Jacques et al., 2014) in a two-step approach. *Step 1*: Run the algorithm for 1–5 clusters using the combined data of each of the MM. Step 2: using the output of *Step 1*, as well as knowledge about the data, the most likely number of clusters was selected for further investigation of clustering strength. We ran the algorithm five times for each MM separately, as well as for the combined data of all MM using this number of clusters, to assess the stability of the outcome. A robust set of parameters was as suggested on page 9–10 of Jacques et al. (2014). Details of the model runs including the parameter sets can be found in the R script in Section SI-1.7.

RQ1b To analyse preference change over time, we selected only those respondents who participated in more than one MM. We quantified preference change through calculation of the Kendall's rank correlation coefficient (τ) (Kendall, 1938) between consecutive MM, as well as between the first and last MM for each respondent. To assess the evolution of preference change over time population-wide, paired student t-tests were performed on τ-values between consecutive MM pairs (i.e., τ_{12} refers to Kendall's τ between the rankings of MM1 and MM2). Thus, all respondents' τ_{12}-values (group 1) were compared to their τ_{23}-values (group 2), etcetera.

RQ1c To explain preference changes over time For CS1_Africa, we used a general linear regression model with beta distribution (Zeileis et al., 2020) to analyse the impact of different independent variables on preference change (dependant variable), measured by Kendall's τ between rankings of consecutive MM (e.g., τ_{12}, τ_{23}). Because the general linear regression model with beta distribution can only be applied to dependant variables ranging between 0 and 1, transformation of the τ-values was required (τ-values can range between −1 and 1). We chose to transform τ-values by making them absolute, thus interpreting (very) negative values as erroneous interpretation of the ranking method through reversal of the scale by the respondent. The impact of such errors on the final outcome is limited, as the number of negative τ-values never exceeded one in 20. We tested the independent variables *language*, *gender*, *age group (low, medium, high)*, *field of work* and *experience (low, medium, high)* using the linear regression model. General linear regression with beta distribution calculates the impact categorical variables value compared to another one (e.g., male to female for gender and low and medium experience to high experience). As both *age group* and *experience* contain three categories, the total number of variables in the model is seven. The analysis was only performed for CS1_Africa, as demographic and experience data was not collected for the two other case studies.

RQ2a To test for convergence of individual preference sets, we first created two sets of rankings of objectives: one containing the rankings of all respondents across all MM and one containing a selection of rankings limited to those respondents who visited all MM resp. 4 out of 5 MM in CS2_CH_Small. We used these sets to create boxplots for each MM, comparing the spread of rankings over time. Furthermore, we calculated Kendall's coefficient of concordance (Kendall & Smith, 1939) to assess the similarity of respondents' rankings per MM.

RQ2b To analyse how aggregated preferences evolve over time, we used the same two sets of objectives ranking data as RQ2a. We then assessed the existence of significant coherent patterns in rankings between respondents using the Friedman test (Friedman, 1937) on these sets for each MM. The Friedman test is a non-parametric statistical test to detect significance of ranking patterns amongst sets of rankings (i.e., answering the question if certain objectives are consistently ranked higher than others). Thereafter, we revealed such patterns through Friedman statistic post-hoc analysis using the R package "agricolae" (de Mendiburu & de Mendiburu, 2019). It provides "aggregate" rankings over all respondents for each MM (Conover, 1998), including information about the significance of the difference between the ranks of the objectives. To assess respondents' learning about the truly important objectives, we followed the ranking of objectives over time visually. We visualised aggregate rankings over time by normalising the rank-sum from the Friedman post-hoc analysis (ordered list with the sums of all individual ranks for the objectives) to the original (but non-discrete) ranking scale between 1 and 10. Next, we analysed the change of these aggregate rankings over time by calculating Kendall's τ-values between MM, analogous to the method from RQ1b.

RQ3 To assess the effect of preference change on MCDA model results, we analysed individual rankings of *alternatives* over time analogous to our analysis of preferences (ranking of objectives) under RQ1b. We calculated these individual rankings of alternatives by performing MCDA (Eisenführ et al., 2010) using aforementioned predictions and value functions. As we elicited objective weights only once, we used the ranking of objectives to approximate weights for each MM and respondent using the *rank sum* (RS) method (Roberts & Goodwin, 2002). RS weights were found to be reliable surrogates for objective weights in MCDA, while being relatively straightforward to calculate, even for higher (i.e. >10) numbers of objectives (Riabacke et al., 2012; Roberts & Goodwin, 2002). We performed MCDA using the ValueDecisions app software (Haag et al., 2022), an online browser-based interface for MCDA that is based on R scripts. As the respondents of all case studies expressed that they did not agree with the axioms the standard additive aggregation model (Lienert et al., 2022), we employed a power-mean aggregation with $\gamma = 0.2$ (Haag et al., 2019b). We ran 200 Monte-Carlo simulations drawing from the uncertainty distributions of the predictions of each attribute. This was the maximum number of simulations possible before running out of memory on the server. We based rankings of alternatives on the descending order of the mean of overall values of each alternative from the Monte-Carlo simulations.

RQ4 To assess whether our findings are comparable across different contexts, for RQ1–3 we systematically compared the outcomes of all case studies to each other, to identify similarities and differences between them.

3. Results

3.1. RQ1a: relationship between individual preferences and explanatory variables

Following our expectations, few significant ($\alpha = 0.05$) results emerged from the ordinal regression between the ranking of objectives and personal characteristics (i.e., explanatory variables) for CS1_Africa (Section SI-2.1.1). Most importantly, we found an effect of the field of work as West African hydrologists assigned a higher rank to the objective *high accuracy of information (11_accur_info)*, but a lower rank to *timely production, distribution and access to information (22_timely_info)* compared to emergency managers (note: all names of objectives and alternatives are given in *Italics*; objectives hierarchy see Figs. SI-1, SI-2 and SI-3). However, contrary to our tentative expectation, we found no significant relationship between experience, measured in years of working with flood management, and the ranking of objectives. Higher age groups assigned a lower rank to *timely production, distribution and access to information*. French-speaking respondents assigned lower ranking to *good support system (43_support_syst)* compared to English-speakers. Interestingly, no statistically significant difference was found between French-speaking and English-speaking respondents for the objective *several languages (23_languages)*, despite the system being available only in English.

The cluster analysis mostly confirmed our expectation that there is no evidence for relationships between latent variables and the ranking of objectives. CS2_CH_Small did not have enough respondents to perform cluster analysis, and no evidence of clustering was found for CS1_Africa and CS3_CH_Larger, with exception of an elicitation method effect (Sections SI-2.1.2 to SI-2.1.4). Indication for the optimal number of clusters K (if any) is given by BIC (indication for clustering) and pi values (where pi = 1 indicates identical ranks within a cluster, and pi = 0 indicates no concordance within a cluster). An "elbow" (lowest point) in the graph with K on the x-axis and BIC on the y-axis, combined with high pi-values signals the presence and number of clusters K. For CS1_Africa and CS3_CH_Larger, increasing the number of clusters K did not result in an elbow in BIC values and did not result in significantly improved pi values. Contrarily, it mostly resulted in additional clusters containing a single or very few respondent rankings. One important exception is MM1 of CS1_Africa, where three ($K = 3$) clear clusters emerged, including one cluster containing identical ranks (pi = 1; i.e., all group discussion members were assigned the same ranking of objectives). These clusters correspond with the three distinct groups that used different ranking elicitation methods in MM1. As no further clustering was observed, we cannot infer evidence for the existence of latent variables driving the ranking of objectives.

3.2. RQ1b: individual preferences over time

The level of preference change varied greatly between respondents (Figs. 2, SI-4 and SI-5), with τ_{mean}-values (mean of τ_{23} and τ_{12}) varying between −0.13 and 0.87 over all case studies.

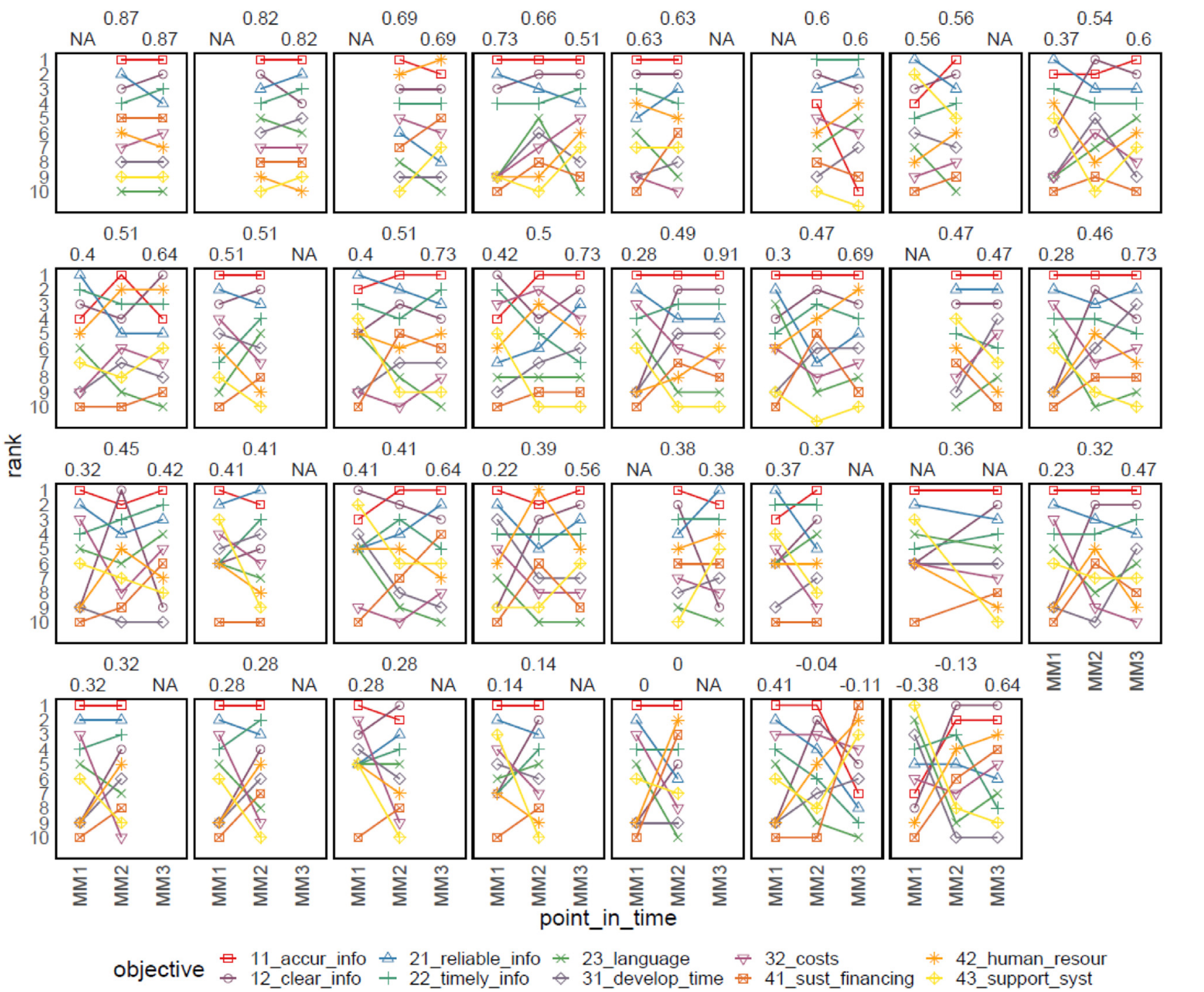

Fig. 2. Preferences over time, measured with the ranking of ten objectives (y-axis) for all 31 individual respondents who visited a minimum of two measuring moments (MM) in CS1_Africa; ordered by decreasing τ_{mean} (mean of τ_{23} and τ_{12}). Individual τ-values between rankings are found above each graph (left: τ_{12}, top middle: τ_{mean}, right: τ_{23}).

While we found the greatest variation in CS1_Africa, this was also the only case study where, for nearly all respondents, preference change decreased over time (higher τ_{23}-values than τ_{12}-values: Fig. 2). This is confirmed by the results of the *t*-test for CS1_Africa, where τ shows a substantial and significant increase of 0.27 ($p = 0.014$). For CS2_CH_Small, we did not observe significant change in τ (i.e. no indication for increasing or decreasing change of preferences, Fig. SI-4), while for CS3_CH_Larger we found a smaller but significant increase of 0.10 ($p = 0.015$) in the beginning (between τ_{23} and τ_{12}), with no significant change thereafter (Fig. SI-5), indicating an initial decrease in preference change between MM1 to MM3, with no change thereafter.

3.3. RQ1c: what influences the likelihood of individual preferences to change over time?

Of the seven independent variables from the demographic and experience variables that we investigated for CS1_Africa, four have a statistically significant impact on preference change between MM1–2 and MM2–3 (Table 2): *gender, age group (low), age group (medium) and experience (low).* However, for the first three of these four variables the impact reverses between MM1–2 and MM2–3.

Table 2
Results of regression model testing the effect of different explanatory variables on preference change (τ between consecutive MM) for CS1_Africa and each MM. *: significant result at $\alpha = 0.05$. **: significant result at $\alpha = 0.01$. MM1–2 pseudo R^2: 0.31. MM2–3 pseudo-R^2: 0.65.

Variable	MM	Estimate	p-value
Language (FR)	MM1–2	0.048	0.593
	MM2–3	0.897**	0.000
Gender (M)	MM1–2	0.286*	0.016
	MM2–3	−1.123**	0.000
Age_group (low)	MM1–2	−0.392*	0.014
	MM2–3	1.332**	0.000
Age_group (med)	MM1–2	−0.682**	0.000
	MM2–3	0.460**	0.001
FOW (HY)	MM1–2	−0.137	0.109
	MM2–3	−0.383**	0.000
Experience (low)	MM1–2	0.464**	0.000
	MM2–3	1.015**	0.000
Experience (med)	MM1–2	0.223*	0.019
	MM2–3	−0.020	0.828

Thus, in line with our expectation, only (low) experience (few years working in flood management) was consistently positively correlated with preference change. The positive estimate indicates that low experience (compared to high experience) was associated with a decrease of preference change (i.e., an increase of τ); in other words, preferences were more stable for respondents with low work experience. Contrary to the results in RQ1a, where field of work was also correlated with preferences (ranking of objectives), here, we found no significant relationship with preference change. There was a correlation between gender and language with preference change, but not in all MM.

3.4. RQ2a: do the individual preference sets converge?

All Kendall's W-estimates of concordance (measure of similarity) were significant ($\alpha = 0.05$; Sections SI-2.3.1 to SI-2.3.3), indicating that the observed patterns are not random. However, contrary to our expectation, concordance between respondents' individual preferences initially increased and then decreased for CS1_Africa (CS1_all stakeholders and CS1_limited_stakeholders), while we observed the opposite trend for the two Swiss case studies (Fig. 3). In line with our working hypothesis, for CS1_Africa and CS2_CH_Small, the concordance between respondents who participated in all MM was higher after MM1 than concordance between the complete set of respondents. For CS3_CH_Larger, the difference in concordance between respondents who participated in all MM and all respondents (irrespective of the number of MM) was already present at MM1, and actually diminished over time. Moreover, concordance varied between objectives (for CS1_Africa with the limited respondent set who had participated in all MM see

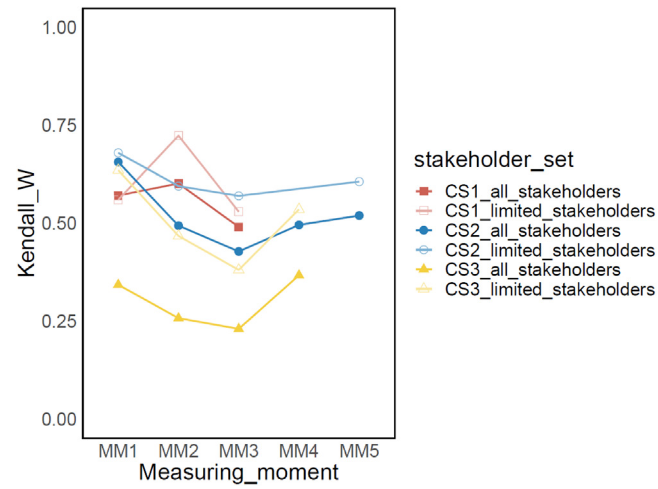

Fig. 3. Concordance (y-axis, Kendall's W) of respondents' ranking of objectives over time (x-axis, MM) for the three case studies (lines). Stakeholder_set: for each case study, the results from all respondents per MM is represented in dark colour (e.g., for CS1_Africa: CS1_all_stakeholders), while a limited set of respondents who visited all MM is represented in light colour (e.g., for CS1_Africa: CS1_limited_stakeholders).

Fig. 4, for CS1_Africa with all respondents see Fig. SI-6, for the other case studies see SI-2.3.5 and SI-2.3.6). Generally, we observed that highly ranked objectives were associated with less spread, and thus more agreement between respondents. This is especially evi-

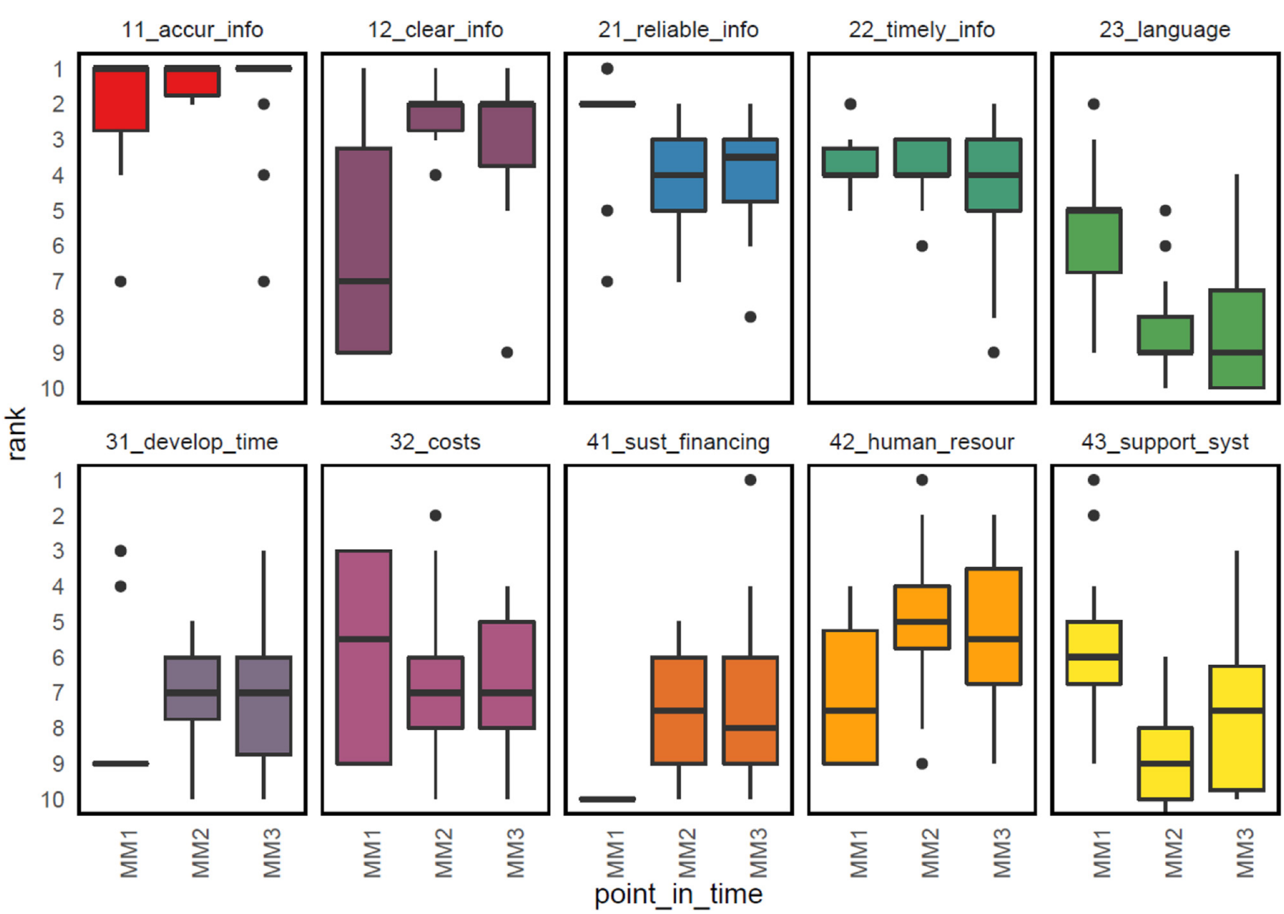

Fig. 4. Respondent rankings of objectives (y-axis, where 1 = best rank, 2 = second best, etc.) over time (x-axis, MM), aggregated in boxplots, separated by ten objectives for CS1_Africa, limited to those respondents who visited all MM. Boxplots indicate the median, 25 and 75 percentiles.

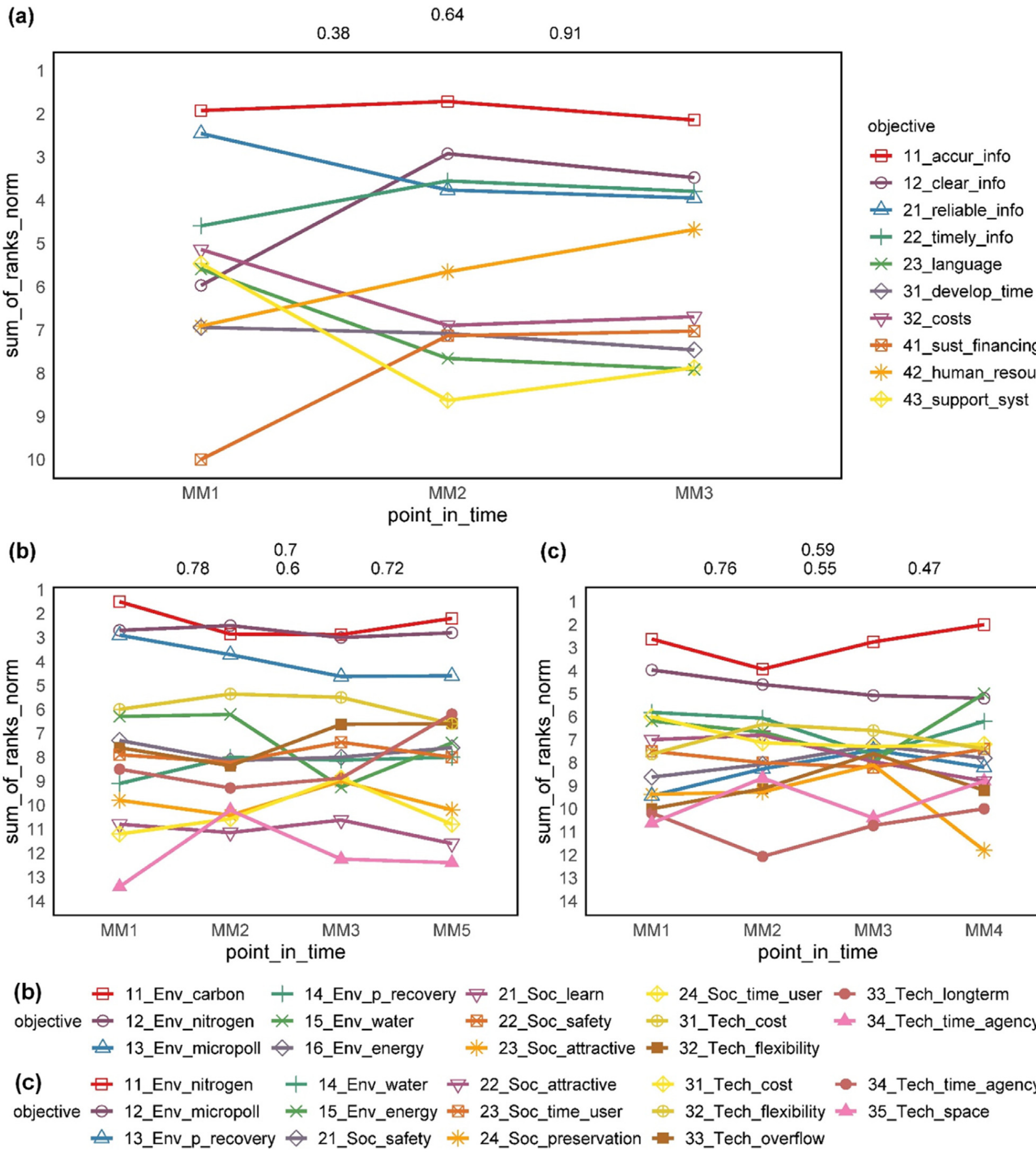

Fig. 5. Aggregated normalised rank-sums of objectives (y-axis) over MM (x-axis) from all respondents, resulting from the Friedman post-hoc analysis, for CS1_Africa, CS2_CH_Small and CS3_CH_Larger (a, b and c respectively). Individual τ-values between rankings are found above each graph (top middle: τ_{mean}, left to right: τ_{12}, τ_{23} and τ_{34} if applicable).

dent for the most important objective *high accuracy of information* in CS1_Africa (*11_accur_info*; Fig. 4).

3.5. RQ2b: how do aggregated preferences evolve over time?

Our tentative expectation was confirmed for all MM and case studies, namely that aggregated rankings seemed to follow coherent patterns and were not random, indicating that certain objectives were consistently ranked higher than others. Specifically, for all MM and in each case study, the post-hoc analysis rankings' p-values were statistically significant (below 0.05). Additionally, in all three case studies, the top ranked objectives remained relatively unchanged through the different MM. On an aggregate level, there was a clear and stable top-1 ranked objective in CS1_Africa (*high accuracy of information*, *11_accur_info*), following the observation for individual respondents in RQ1b (Fig. 2). In the Swiss case studies, the aggregated top-3 ranked objectives were stable in CS2_CH_Small and the top-2 ranked objectives in CS3_CH_Larger (Fig. 5). Contrary to our expectation, we observed different patterns between the case studies: a sharp decrease of changes in the ranking of objectives over time (between MM2 and MM3) for CS1_Africa and the opposite tendency for CS3_CH_Larger, while

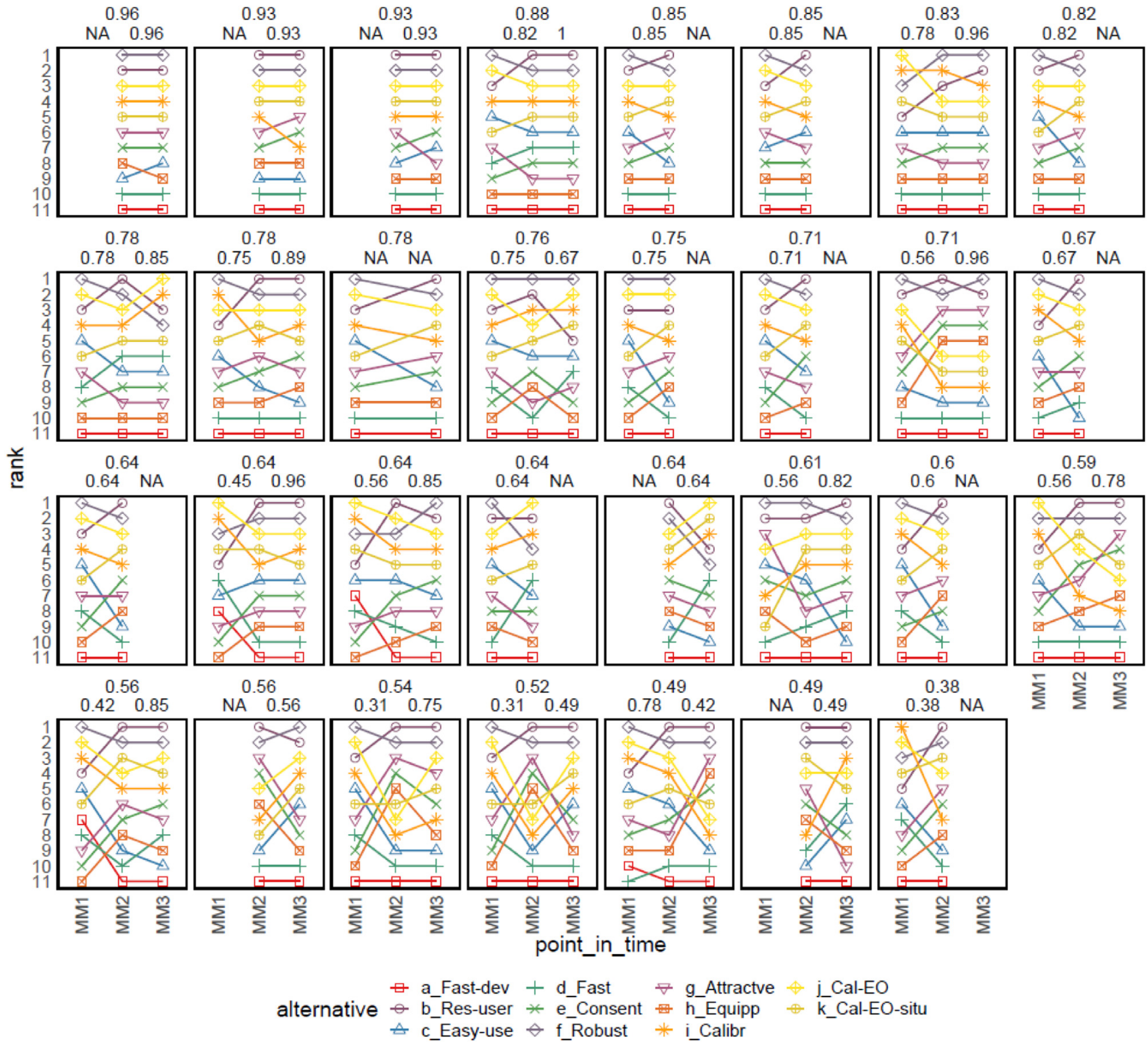

Fig. 6. Individual ranking of alternatives (y-axis) resulting from the MCDA over time (MM; x-axis) using RS weights from individual ranking of objectives over time, for all respondents (boxes) who visited a minimum of two MM in CS1_Africa. Ordered by decreasing τ_{mean} (mean of τ-values of rankings between MM1-MM2 and MM2-MM3 etc.). Individual τ-values between rankings are found above each graph where relevant (left: τ_{12}, top middle: τ_{mean}, right: τ_{23}). Rank 1 = best, rank 11 = worst.

CS2_CH_Small showed no clear pattern (Fig. 5). Boxplots representing each objective's ranking over time are presented in Fig. 4 and Section SI-2.3.4–SI-2.3.6. Aggregated preference change of the selection of respondents who participated in all MM were similar to those of the complete set of respondents, evident from the similarity between τ-values (Table SI-4).

3.6. RQ3: what is the impact of changing preferences on the outcome of an MCDA on an individual and group level in three real-world case studies?

As expected, the top-ranked alternatives were relatively similar across respondents in all case studies (Figs. 6, SI-11 and SI-12), indicating their robustness both through time and between respondents. Also as expected, individual ranking of alternatives over time displayed lower change than the individual ranking of objectives (compare Fig. 6 to Fig. 2). This was not only the case for CS1_Africa, but also for the two Swiss case studies (comparing Figs. SI-11 to SI-4 and Figs.SI-12 to SI-5). Again as expected, and similar to individual ranking of alternatives, aggregated rankings over time in all but one instance (τ_{23} of CS1_Africa) resulted in higher τ-values than for the ranking of objectives over time, indicating comparatively less change of alternatives (Fig. 7). All case studies displayed a clear and stable top section of one or two highly ranked alternatives (following our working hypothesis), an unstable middle section and a clear and stable bottom section of one or two lowly ranked alternatives (which we had not anticipated; Fig. 7). Changes in ranking of alternatives for CS1_Africa followed a similar pattern to that emerging from the rankings of objectives underlying these ranking of alternatives, albeit not fully reflected in the τ-values (see results RQ2b, Fig. 7). This means that individual ranking of alternatives tended to stabilise over time, as reflected by the estimate value (mean of differences between τ_{12} and τ_{23}), which equalled -0.21, p-value = 0.005. This is a very significant stabilisation, considering that the τ-values vary between 0 and 1. For CS2_CH_Small and CS3_CH_Larger, the translation of the patterns of change was less clear (compare Fig. 5 to Fig. 7).

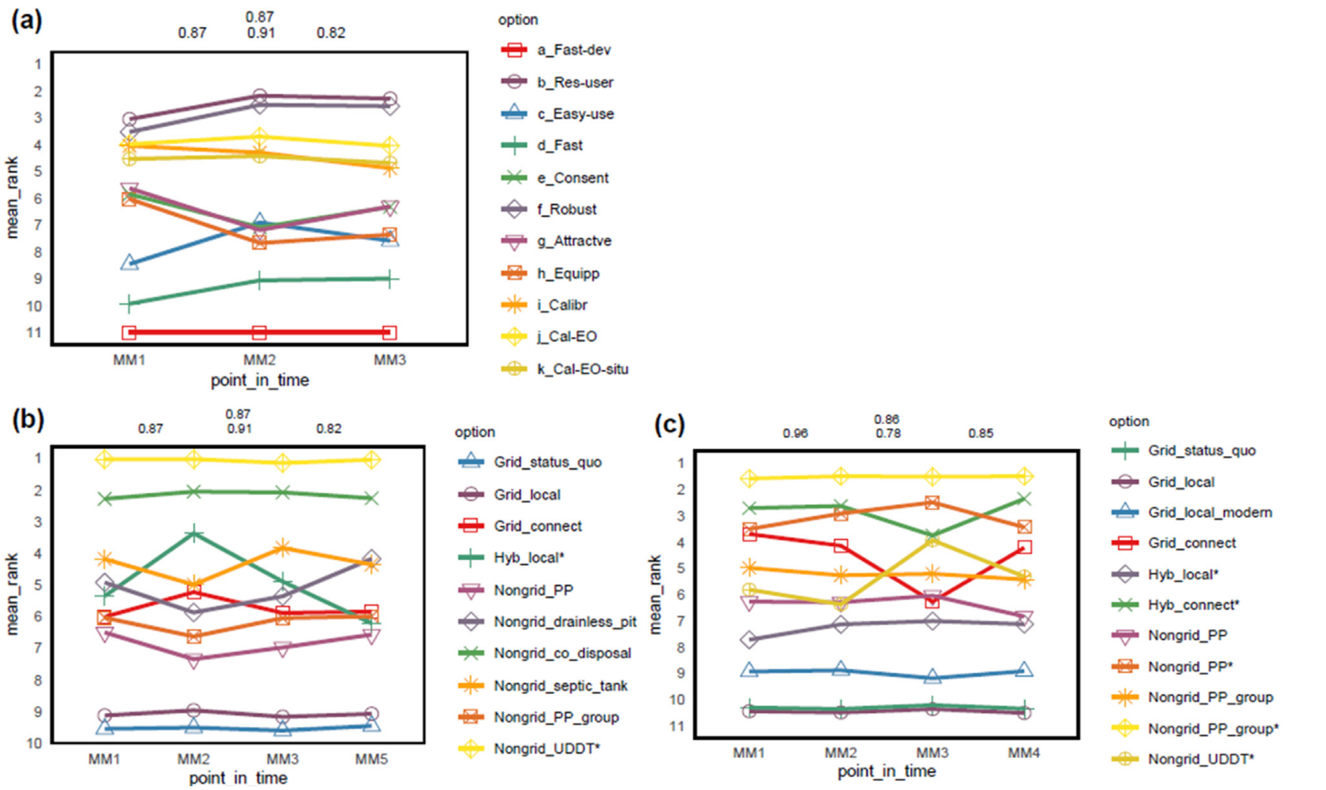

Fig. 7. Aggregated mean rankings of the alternatives (y-axis) resulting from the MCDA over time (MM, x-axis) using RS weights from rankings of objectives, with τ-values of changes between MM (facet title bottom, left to right) and the overall τ_{mean} (facet title top centre) for CS1_Africa, CS2_CH_Small and CS3_CH_Larger (a, b and c respectively). *Alternatives that include source separation.

3.7. RQ4: is there evidence that the findings from RQ1–3 apply across case studies?

While the majority of the findings were identical between the case studies, there were also some important differences (Table 3). Most notably, the individual preferences evolved distinctively different over time between cases, where preference changes in CS1_Africa decreased, while this trend was not or barely present in the other two case studies. Aggregated preferences reflected these trends. These patterns further translated to the aggregated ranking of alternatives for CS1_Africa, but not so clearly to the other two case studies. Finally, concordance between individual rankings of CS1_Africa increased first and then decreased, an opposite trend to the other case studies.

4. Discussion

We discuss the results of each research question separately and in detail in SI-3. Here, we discuss the overall findings and end with a discussion of our assumptions, limitations and avenues for further research.

4.1. Preference change and decision-making across contexts

As discussed above, preference change was universally observed across our case studies. We found no significant predictor for preferences, while the only consistent and significant predictor for preference change in CS1_Africa was experience with flood management: Low experience was associated with lower preference change. This indicates that for respondents that were relatively new to the field, preferences were influenced less easily by repeated exposure to the FANFAR FEWS and discussions during the workshops. Although the possibility of randomness should not be discounted, this finding counters our expectation that pre-existing preferences become more clear (or rigid) with experience. It also contrasts earlier findings, where experience was completely unrelated with preference stability over two measuring moments (Lienert et al., 2016) and contrasts with literature expectations that experienced decision makers may have existing, retrievable preferences (see RQ1a; e.g., Bettman et al., 2008; Warren et al., 2011).

In the absence of any other clear impact of demographic or experience variables (RQ1a; RQ1c), we suggest that preferences could change because (a) respondents better understand the decision problem, e.g., thanks to increasing information over time (factual learning) and/or having direct experience with the alternatives; (b) respondents better understand their own preferences (preference learning, see Aubert & Lienert, 2019) and/or the salience of objectives increases (especially if they are not achieved); (c) influences of specific interventions in the decision-making process (e.g., method effects; see Franco et al. (2021) and references therein); and (d) have interacted with each other and better understand the values and preferences of other respondents (group decision-making, but possibly including occurrence of biases, see Kerr & Tindale, 2004; Montibeller & Von Winterfeldt, 2015, 2018). While preferences differ from the underlying attitudes, the central route proposed to drive attitude change would be followed for our suggested explanations a, b and d, while c (intervention effects) would follow the peripheral route (Section 1.2). Preference change and congruence between stakeholder preferences (mental consensus building) has been hypothesised to result from collaborative modelling exercises similar to the practices in our workshop. This hypothesis was only recently confirmed by de Gooyert et al. (2022) using empirical data from eight case studies in The Netherlands.

Table 3
Summary and comparison of findings per case study (CS). CS1: CS1_Africa, CS2: CS2_CH_Small, CS3: CS3_CH_Larger. Findings that were identical between case studies are indicated with: "=". Findings indicated with "≈" were similar to each other and to those indicated with "=". Findings indicated with "≠" were different from each other and from those indicated with "=" and "≈". Findings that confirm our hypotheses are indicated in bold and green, findings that contradict our hypotheses were indicated in regular typeface.

RQ	CS1	CS2	CS3	Details
1a	**=**	**=**	**=**	We did not find a relationship between latent variables and ranking of objectives in the cluster analysis in any CS, with exception of a method effect (three different elicitation methods) in MM1 in CS1_Africa. Ordinal regression to find relationships between explanatory variables and ranking of objectives was only conducted for CS1_Africa, where we found hardly any effects.
1b	**≈**	≠	**≈**	For all CS, preference change varied per individual. While preference change decreased for CS1_Africa and to a lesser extent in CS3_CH_Larger, CS2_CH_Small did not exhibit such a trend.
1c	NA	NA	NA	The relationship of demographic and experience variables were only explored for CS1_Africa, where we only found consistent effects from low experience, but in opposite direction than expected from previous research results in literature.
2a	≠	≈	≈	The trend between CS1_Africa and the Swiss CS are opposite: in the former, concordance increased first before decreasing later. In the latter, concordance first decreased before increasing. For all CS, concordance at the final MM was lower than at the first MM.
	=	**=**	≠	For CS1_Africa and CS2_CH_Small, respondents who visited all MM had higher concordance with each other after the first MM than respondents who did not. This was not the case for CS3_CH_Larger.
	=	**=**	**=**	Respondents in all CS showed more agreement (less spread in ranking) on the highest ranked objectives compared to the other objectives.
2b	**=**	**=**	**=**	Aggregated ranking of objectives followed coherent patterns and were significant (not random) for all CS.
	≈	**≈**	**≈**	The top ranked objectives of all case studies were relatively stable over time (in all MM), although this was true for different numbers of top-ranked objectives per CS, ranging from 1 to 3 objectives.
	≠	≠	≠	Aggregated preference change patterns were different for each CS. For CS1_Africa changes diminished over time (i.e., preference stability increased), for CS2_CH_Small changes fluctuated and for CS3_CH_Larger changes increased over time (i.e., stability decreased).
	=	=	=	Aggregated preference changes of respondents who visited all MM were similar to those of the complete set of respondents in all CS.
3	**=**	**=**	**=**	Top-ranked alternatives were very similar across respondents within each CS.
	=	**=**	**=**	Rankings of alternatives on both individual and group level changed less than rankings of objectives in all CS (rankings of alternatives are more stable than rankings of objectives).
	=	**=**	**=**	Group rankings of alternatives had stable top and bottom-ranked sections, but a middle section that changed over time in all CS.
	≠	≈	≈	Group rankings of alternatives for CS1_Africa showed a similar pattern to the ranking of objectives underlying them, which was not the case for CS2_CH_Small and CS3_CH_Larger.

The preference changes we observed translated to the aggregated group level (RQ2a, RQ2b), suggesting their potential impact on group decision-making. However, the impact on the performance of alternatives was much less pronounced and on an aggregated level, the best performing alternative was always stable (RQ3). The performance of alternatives have been shown to be relatively insensitive to preference changes in previous studies that performed uncertainty and sensitivity analyse (e.g., Haag et al., 2019b; Lienert et al., 2022, 2016). This was also found in sensitivity analyses of the Swiss and African case studies in this paper, using the full set of preference parameters elicited from stakeholders with standard methods (Beutler et al., 2021; Lienert et al., 2022). Observed stability on aggregate and alternative levels also results from averaging and the elimination of noise. Not only does this provide decision makers with the necessary confidence about the robustness of a decision, but our results based on fast and frugal elicitation procedures also questions the indispensability of resource intensive, sophisticated preference elicitation methods. To date, it seems unclear whether such faster elicitation methods perform sufficiently well to inform large strategic decisions of the type presented here, an issue that has recently been addressed by Katsikopoulos et al. (2018). There is an urgent need for further research in this area.

It is encouraging to see that there were no consistent differences between results from the African case study and the Swiss case studies. This suggests that no systematic bias occurs when using elicitation methods across different cultural contexts. Differently put: the construction and evolution of personal preferences in public decision-making seems to follow similar mechanisms across cultures and application cases. We are not aware of literature; and specific research comparing MCDA methods across cultural contexts is needed to confirm our observations. One exception was the pattern of concordance between preferences, which followed opposite trends between the African and Swiss case studies. However, this (and possibly more) observations could perhaps be explained by the different decision context and particulars of each case: while respondents from the Swiss cases were restricted to imagining what the alternatives could be like until after implementation sometime in future, respondents in CS1_Africa were confronted with the actual alternatives throughout the project period. Such individual differences between case studies thus may change the way individual preferences of participants towards objectives evolve. For instance in the African case, the (flood forecast and alert system) alternative was continuously improved, based on the participants' preferences, and objectives that were perceived as covered might have received a lower ranking than before. In contrast, presenting unexpected MCDA results about best-performing alternatives seems to have had a disruptive effect on preferences in the Swiss cases. Here, learning that undesired alternatives performed best may have led to preference shifts and/or (unconscious) biases because the stakeholders aimed to align their preferences with what they perceived as their favoured alternative (discussed in SI-3.1.5).

4.2. Assumptions, limitations and further research

Empirical evidence on preferences from real-world case studies is tied to a number of assumptions and limitations. Firstly, the fast elicitation of preferences applied in our study is not recommended for use in an MCDA process (Eisenführ et al., 2010). French (2021) recently argued that part of the purpose of decision analysis is to help stakeholders contextualise their preferences and values, making their change inherently part of such a process. This sets preferences apart from the elicitation of parameters or probabilities, which are used to measure the outside world. This distinction poses an important challenge of calibration and validation to any user of a model that relies on qualitative judgement. While parameters and probabilities can be determined in the presence of appropriate methodologies and instrumentation, we lack such methods to calibrate preferences. We can thus not be sure that the tools we used captured what we intended to capture. Perhaps respondents did not fully understand the task, as was observed in online elicitation of weights (Aubert et al., 2020; Lienert et al., 2016). Another possibility is that respondents were unable to accurately represent their preferences on the provided measurement scale. While certain tendencies were observed across case studies (e.g., the presence of preference change), inconsistencies might be at least partly explained by the randomness arising from the measurement method. Furthermore, seemingly random variations likely occur as preferences are volatile and context dependant reflections of more stable underlying attitudes (Eagly & Chaiken, 2007). For all case studies described in this paper, we have executed a full MCDA using more robust elicitation methods, including Swing (see Lienert et al. (2022) for CS1_Africa, and Beutler et al. (2021) for the two Swiss case studies). Such methods put considerable burden on respondents and researchers in terms of time and cognitive effort, and are thus not suitable for repeated application over time with the same respondents. As RS weights have been shown to perform reasonably well to present preferences (Roberts & Goodwin, 2002), we regard this method as sufficient for the identification of preference change. However, future research would benefit from direct comparison of the fast and frugal and well-known standard preference elicitation methods, possibly including other preference parameters such as the shape of marginal value functions, or even the aggregation model.

Secondly, we compared case studies of different size, timing and context, using semi-standardised approaches that are less controlled than an experimental set up. While this increases the uncertainty of the findings, it allows for interesting insights in the generalisability of findings across different cases. To complement experimental studies and verify the practical validity of their findings on individual preference construction and change, empirical studies are essential (e.g., Franco et al., 2021). As reality is often messy, some methodological inconsistencies between study cases are inevitable. Further research should focus on attempting to reproduce our findings in other decision contexts, as empirical evidence is largely lacking. For example, research could focus on systematic comparison of elicitation methods and associated mechanisms of preference construction and representation across different cultures. For future studies, we recommend to invest in lasting relationships and commitment from study participants, to ensure respondent continuity over time.

We identify some important avenues for future research. Firstly, the abovementioned limitations of preference measurement provide an important open research question. French (2021) suggests that the literature on metacognition could provide a starting point. Secondly, future research should focus on reproducing the findings that appear consistent between our case studies in other real-world decision contexts, to verify e.g., that (i) preference construction and preference change cannot be explained by personal variables (such as demographics) or latent variables, (ii) top ranked objectives are more stable and more agreed upon, (iii) alternatives from an MCDA using elicited preferences change less than the preferences themselves and (iv) the top-ranked alternatives are widely agreed upon and very stable over time. Thirdly and more specifically, a better understanding of group decision-making processes and possible group biases (including de-biasing techniques) is urgently needed (e.g., Kerr & Tindale, 2004; Montibeller & Von Winterfeldt, 2015, 2018). For example, the "confirmation bias" and related "desirability of options/choice bias", potentially observed in our case studies, should be further investigated. Given the importance of group workshops in MCDA processes, there seems to be an astonishing lack of specific research in this field. Fourthly, research should attempt to uncover (i) what makes preferences change (e.g., learning, interventions, interactions), (ii) why in certain cases preference change seems to diminish over time, while in others it does not and (iii) why preference alignment between stakeholders sometimes increases but in other cases decreases over time. This latter research avenue follows a "process perspective", which should include observational research with the aim of uncovering the development of cognitive processes and narratives of individuals and groups over time (Franco et al., 2021). In this regard, we could draw learnings from social psychology theories, including the persuasion/ attitude change literature to better understand the underlying processes of our interventions. Finally, we should explore whether the wide agreement on, and stability of alternatives we found indicates that fast preference elicitation might suffice for some decision cases, for example by comparing the MCDA outcomes for fast elicited preferences with those acquired using robust methods for weight elicitation such as Swing or Trade-off. We also propose that interactive, flexible elicitation deserves more attention. This could include eliciting some (fast and frugal) preference data from stakeholders (e.g., de Almeida et al., 2016) and making rough predictions for outcomes of alternatives, followed by preliminary sensitivity analyses. Results could then inform where next activities are best invested: making more precise predictions or using better methods to elicit stakeholder preferences.

5. Conclusion

Large scale public projects require robust, consensus decision-making. Our study sheds new light on the construction and change of preferences of stakeholders in complex, real-world public decision-making. While former studies mainly focussed on preference construction and change in controlled contexts, our findings confirmed that also in real-world and complex contexts, we find some universal mechanisms. Preference construction is barely impacted by demographic or other personal or circumstantial variables and even unknown, or latent variables do not seem to play a role across our diverse contexts. Thus, the construction of preferences cannot be explained, or predicted for stakeholders. Although preferences vary between individuals, we observed wide agreement on the most important objectives in our diverse decision-contexts.

While individual preferences vary between stakeholders, they also change significantly over time. Similar to preference construction, we found little evidence for personal or external variables to explain the observed changes. We found some evidence for a reduction of such changes over time both on an individual and group level, but this evidence was not consistent over our decision contexts. We found different and opposing patterns of evolution in the agreement of stakeholders between our African and Swiss case studies, signs of learning and the forming of consensus were observed between those stakeholders that were repeatedly involved over time. Despite the preference changes, the most important objectives remained stable over time and agreed upon by stakeholders.

Public decision-making processes should aim at the selection of a policy or investment alternative that performs “best” in meeting the objectives of a large diverse set of stakeholders. We used MCDA following MAVT to establish and rank alternatives for our case studies, using the diverse and changing stakeholder preferences as decision inputs. The observed differences in preferences between stakeholders as well as their changes were reflected in the rankings of alternatives, but greatly diminished. Indeed, the consistent (in time) agreement between stakeholders on the most important objectives resulted in the same top-ranked and bottom-ranked alternatives over time across stakeholders in all case studies. Thus, despite the diverse and changing preferences of stakeholders in complex public projects, we can arrive at an important conclusion, critical for those involved in large-scale public decision-making: it is possible to arrive at robust consensus-based decisions by selecting stable, well performing alternatives. Although the conclusions from this study were drawn using evidence from three distinctive case studies, we are only at the beginning of understanding the mechanisms and consequences of changing preferences in public decision-making. This study is one of the first of its kind, and our findings should thus be verified across decision-making contexts and projects.

Data availability

The data and R analysis scripts are available in the Eawag Research Data Institutional Collection (ERIC open) via the following DOI:10.25678/0006PX.

CRediT authorship contribution statement

M. Kuller: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Validation, Writing – original draft, Writing – review & editing, Visualization, Project administration. **P. Beutler:** Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Validation, Writing – original draft, Writing – review & editing, Visualization, Project administration. **J. Lienert:** Conceptualization, Methodology, Investigation, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Acknowledgments

We warmly thank the FANFAR consortium (CS1_Africa) and the research team at Eawag (CS2_CH_Small and CS3_CH_Larger) for organising the different case studies, and the participants of all three case studies for their repeated and committed collaboration. We thank Francisco Silva Pinto and Daniel Hofmann for their work on data collection and data entry. We furthermore thank Andreas Scheidegger for his continued support with the design of our data analysis methodology. Finally, we thank our two anonymous reviewers for their valuable contributions and corrections. This work was supported by the European Union’s Horizon 2020 research and innovation programme under grant agreement No 780118 and by the Canton of Solothurn (Switzerland). This research was further supported by the Eawag Discretionary Funding scheme to offset impacts of COVID-19.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ejor.2022.12.001.

Appendix A: working hypotheses

In this study, we followed an exploratory approach with relatively open research questions. However, based on own experience and literature, we do have some expectations regarding results. We formulated our expectations as working hypotheses, including explanations (summary see Table 1).

RQ 1a What is the relationship between the dependant variable individual stakeholder preferences, and independent variables demo-graphic variables, personal experience and unknown (latent) variables?

Working hypotheses 1a. We expect no or weak relation between personal, demographic variables, and preferences, based on own results from surveys (e.g., Lienert et al., 2016) and confirmed by a recent review (Franco et al., 2021). The latter also found little evidence for an effect of personal experience on behaviour, but state that there is a lack of research. Psychological literature indicates that experienced decision makers may be able to retrieve existing preferences and thus be less susceptible to framing effects (see Section 1.1 and e.g., Bettman et al., 2008; Warren et al., 2011). Since preferences would have been previously formed during experiences with flood- and wastewater management over the years, they could be more stable over the different MM (also see below; RQ1c), and less influenced by other variables. This tentative working hypothesis, however, was not supported for experience by our earlier survey, whereas knowledge about wastewater management was significantly correlated with preference stability (Lienert et al., 2016). We have no expectations regarding latent variables.

RQ 1b Do we observe significant patterns of change in preference rankings of objectives for individual stakeholders over time with interventions?

Working hypotheses 1b. We expect preferences of individuals to change over time, because preferences are constructed during decision-making processes and are susceptible to various influences (as reviewed in Section 1.1 and e.g., Lichtenstein & Slovic, 2006). Providing detailed information on objectives, attributes and alternatives' performances during our interventions, we expect a detectable change in preferences. We also hypothesise that an individual learns more about their own values and preferences. Specifically, if preferences about the importance of objectives are formed during the first workshop (MM1), and these objectives are repeatedly reflected upon during consecutive workshops by individual participants, the preference construction process might be solidified, and preferences concerning the importance of objectives will be clearer and easier to retrieve from memory as time goes on. We therefore additionally hypothesise that the variance and fluctuation of an individuals' preferences decrease over time. Increasing stability of individual preferences over time is supported by test-retest experiments from economics (briefly reviewed in Section 1.2). Moreover, there may be an interplay with group decision-making effects and group biases. For instance, the well-known groupthink bias (see also RQ 2a below, Janis, 1971) could in our examples lead to preferences aligning and becoming more stable over time, also for individuals (for reviews see e.g., Kerr & Tindale, 2004; Montibeller & von Winterfeldt, 2018).

RQ 1c What is the relationship between the dependant variable “likelihood of a change in individual stakeholder preferences” (if any) and the static independent demographic variables, personal experiences, and unknown (latent) variables?

Working hypothesis 1c: Since we do not expect a strong influence of explanatory variables on preferences (RQ1a), we would also not expect a strong relationship between these variables and the evolution of preferences over time in CS1_Africa (data not available for the two Swiss cases). This was the main result in our earlier survey of preference stability (objectives weights, elicited online twice; Lienert et al., 2016). A possible exception could be work experience (see RQ1a), but it would contrast our earlier results, where we found no effect of experience (Lienert et al., 2016). Knowledge could be another exception, but we did not elicit this variable separately in the studies presented in this paper.

RQ 2a Do we observe convergence between stakeholder preferences (i.e., more agreement between stakeholders)? Does convergence differ between respondents who attended all workshops (i.e., MM), and those who did not?

Working hypotheses 2a: Due to group processes and biases occurring in groups, we tentatively expect that preferences of individuals converge. Kerr & Tindale (2004, pp. 632–633) state: “Group decision-making research in the 1960s and 1970s typically emphasised the processes involved in moving from a diverse set of individual positions or preferences to agreement on a consensus choice for the group. […] However, the dominant paradigm behind recent group decision-making research has focused on information rather than on preferences”. Recently, Norström et al. (2020), p. 188) described the need to analyse such processes as: “assessing the [interactive] principle should also focus on capturing learning, how the perceptions of actors change throughout the process, and the degree to which a shared perspective emerges”. Convergence in groups may be caused by biases such as groupthink, which postulates that group members focus on achieving consensus without realistic appraisal of alternative courses of action (Janis, 1971). This bias, however, seems more relevant if alternatives are directly discussed, while value-focused MCDA focusing on preferences for objectives might partly overcome this problem. Regarding preferences, two biases might be more relevant: Group polarisation states that group discussions enhance the opinion that was initially held by the majority. False consensus implies that individuals overestimate the similarity of their own preferences with that of others and anchoring on the judgement of others occurs (see Montibeller & von Winterfeldt (2018) and references therein; also see RQ1b above and e.g., Kerr & Tindale (2004)). Given the importance of group processes for decision analysis, there is a surprising lack of recent research (Montibeller & Von Winterfeldt, 2015, 2018), and we can only formulate a very tentative expectation. Additionally, if such convergence effects occur, we expect them to be stronger for individuals that attended more workshops.

RQ 2b How do aggregated (group) preferences evolve over time? Is there a difference between the subset of respondents who attended all MM of their case study, compared to the entire set of respondents?

Working hypothesis 2b: First, we tentatively expect that aggregated rankings over all individuals in a group follow coherent patterns and are not random, or in other words that some objectives are consistently ranked higher by the group as whole. This is based on the idea that certain objectives are generally more important in flood forecasting and alerts (e.g., receiving accurate information well before a flood) or wastewater management (e.g., environmental and human health protection) than others (e.g., low costs or high user comfort). An additional working hypothesis is that group phenomena might lead to the convergence of stakeholders’ preferences, introduced above (RQ2a). Therefore, we might expect that the preferences of the group as a whole are reinforced in each workshop and are thus less influenced by other variables. Rather, something like a “group memory” or “group worldview” may emerge over time, leading to more stable aggregated group preferences. This is caused by several rounds of interventions (similar to a group Delphi process: Rowe & Wright, 1999), where individuals have been repeatedly thinking through the decision on their own as well as discussing it with peers. They may have learnt more about the decision through our interventions, e.g., experimenting with the FEWS in each FANFAR workshop in CS1_Africa or receiving fact sheets about each alternatives’ pros and cons in CS2_CH_Larger. Our tentative hypothesis is thus that there are reduced changes of aggregated group preferences over time. Furthermore, we expect to find some evidence for collective learning of truly important objectives. Additionally, we expect the reduction in aggregated changes over time to be larger for respondents that visited all workshops (see RQ2a).

RQ 3 What is the impact of changing preferences on the outcome of an MCDA (i.e., ranking and overall value of alternatives) at an individual and group level in three real-world case studies?

Working hypothesis 3: Given that the above hypotheses hold, we tentatively postulate that the best-performing alternatives are similar for all respondents within a case study. Additionally, we expect that stabilised preferences of individuals and the entire group over time (MM) will translate in more stable MCDA results over time, i.e., clearer, more consistent rankings of alternatives. Furthermore, we expect higher stability of the alternatives compared to objectives, because the additional MCDA processing step levels out differences in preference parameters. We have found such a “mitigating” effect in earlier studies, where it was possible to find compromise alternatives over the entire group, despite individual differences concerning the importance of objectives (e.g., Haag et al., 2019b; Lienert et al., 2016). Moreover, good performance of alternatives is based on achieving those objectives that are most important to stakeholders. We think that these most important objectives are more salient to the stakeholders and are thus more stable over time (the MM), compared to less important and less salient objectives. This would translate into the best-performing alternatives also being more stable over time compared to lower-ranked ones. Finally, we expect the change patterns of aggregated (group) rankings of alternatives to be reflected in the patterns of aggregated ranking of objectives (RQ2b), since objectives are an important underlying dataset for the MCDA.

RQ 4 How do the answers to the above questions compare between case studies? Do we find evidence for universality of findings between different decision contexts?

Working hypothesis 4: Can we expect that preference construction processes for public policy decisions are similar in the human mind, irrespective of the context such as the region (Global North or South), the application (flooding or wastewater management) and the exact decision case? Given that this highly explorative expectation holds, we expect to receive similar results and patterns across all case studies. On the other hand, if preference construction in such public policy decisions is not a universal process, i.e., if this working hypothesis is rejected, we might expect the two cases from Switzerland CS2_CH_Small, CS3_CH_Larger) to be rather similar because they concern a very similar application case from the same country. In contrast, we might observe a larger difference between both Swiss cases and the FANFAR case in West Africa (CS1_Africa). We are not aware of specific research addressing such questions.

References

Anderson, R. M., & Clemen, R. (2013). Toward an improved methodology to construct and reconcile decision analytic preference judgments. *Decision Analysis, 10*(2), 121–134.

Andersson, J., Ali, A., Arheimer, B., Crochemore, L., Gbobaniyi, B., Gustafsson, D. et al. (2020,. 4-8 May). Flood forecasting and alerts in west africa-experiences from co-developing a pre-operational system at regional scale. Paper presented at the EGU General Assembly, Vienna, Austria. doi:10.5194/egusphere-egu2020-7660.

Angie, A. D., Connelly, S., Waples, E. P., & Kligyte, V. (2011). The influence of discrete emotions on judgement and decision-making: A meta-analytic review. *Cognition & Emotion, 25*(8), 1393–1422.

Ariely, D., Loewenstein, G., & Prelec, D. (2003). Coherent arbitrariness": Stable demand curves without stable preferences. *The Quarterly Journal of Economics, 118*(1), 73–106.

Aubert, A. H., & Esculier, F. (2020). Recommendations for online elicitation of swing weights from citizens in environmental decision-making. *Operations Research Perspectives, 7*, Article 100156. https://doi.org/10.1016/j.orp.2020.100156.

Aubert, A. H., & Lienert, J. (2019). Gamified online survey to elicit citizens' preferences and enhance learning for environmental decisions. *Environmental Modelling & Software, 111*, 1–12. https://doi.org/10.1016/j.envsoft.2018.09.013.

Belton, V., & Stewart, T. J. (2002). *Multiple criteria decision analysis: An integrated approach*. Springer Science & Business Media.

Berrens, R. P., Bohara, A. K., Silva, C. L., Brookshire, D., & McKee, M. (2000). Contingent values for New Mexico instream flows: With tests of scope, group-size reminder and temporal reliability. *Journal of Environmental Management, 58*(1), 73–90. https://doi.org/10.1006/jema.1999.0308.

Bettman, J. R., Luce, M. F., & Payne, J. W. (1998). Constructive consumer choice processes. *Journal of Consumer Research, 25*(3), 187–217.

Bettman, J. R., Luce, M. F., & Payne, J. W. (2008). Preference construction and preference stability: Putting the pillow to rest. *Journal of Consumer Psychology, 18*(3), 170–174.

Beutler, P., Larsen, T.A., .Maurer, M., Staufer, P., & Lienert, J. (2021). Decision-making for the transition from central to novel wastewater infrastructure systems. Project website in-cluding link to open access articles. Retrieved from https://www.eawag.ch/en/department/ess/projects/decision-making-for-wastewater-infrastructure-transitions/on 04 August 2021.

Biernacki, C., & Jacques, J. (2013). A generative model for rank data based on insertion sort algorithm. *Computational Statistics & Data Analysis, 58*, 162–176.

Bliem, M., Getzner, M., & Rodiga-Laßnig, P. (2012). Temporal stability of individual preferences for river restoration in Austria using a choice experiment. *Journal of Environmental Management, 103*, 65–73. https://doi.org/10.1016/j.jenvman.2012.02.029.

Brouwer, R. O. Y., Akter, S., Brander, L., & Haque, E. (2009). Economic valuation of flood risk exposure and reduction in a severely flood prone developing country. *Environment and Development Economics, 14*(3), 397–417 10.1017/S1355770×08004828.

Brouwer, R. (2006). Do stated preference methods stand the test of time? A test of the stability of contingent values and models for health risks when facing an extreme event. *Ecological Economics, 60*(2), 399–406. https://doi.org/10.1016/j.ecolecon.2006.04.001.

Brouwer, R. (2012). Constructed preference stability: A test–retest. *Journal of Environmental Economics and Policy, 1*(1), 70–84. https://doi.org/10.1080/21606544.2011.644922.

Brouwer, R., Logar, I., & Sheremet, O. (2017). Choice consistency and preference stability in test-retests of discrete choice experiment and open-ended willingness to pay elicitation formats. *Environmental and Resource Economics, 68*(3), 729–751. https://doi.org/10.1007/s10640-016-0045-z.

Cameron, J. I. (1997). Applying socio-ecological economics: A case study of contingent valuation and integrated catchment management. *Ecological Economics, 23*(2), 155–165. https://doi.org/10.1016/S0921-8009(97)00051-7.

Carson, R. T., Hanemann, W. M., Kopp, R. J., Krosnick, J. A., Mitchell, R. C., Presser, S., et al., (1997). Temporal reliability of estimates from contingent valuation. *Land Economics, 73*(2), 151–163. https://doi.org/10.2307/3147279.

Conover, W. J. (1998). *Practical nonparametric statistics*. John Wiley & Sons (Vol. 350).

de Almeida, A. T., de Almeida, J. A., Costa, A. P. C. S., & de Almeida-Filho, A. T. (2016). A new method for elicitation of criteria weights in additive models: Flexible and interactive tradeoff. *European Journal of Operational Research, 250*(1), 179–191. https://doi.org/10.1016/j.ejor.2015.08.058.

de Gooyert, V., Rouwette, E., van Kranenburg, H., Freeman, E., & van Breen, H. (2022). Cognitive change and consensus forming in facilitated modelling: A comparison of experienced and observed outcomes. *European Journal of Operational Research, 299*(2), 589–599. https://doi.org/10.1016/j.ejor.2021.09.007.

de Mendiburu, F., & de Mendiburu, M.F. (.2019). Package 'agricolae'. R Package, Version, 1.2-1.

Eagly, A. H., & Chaiken, S. (2007). The advantages of an inclusive definition of attitude. *Social Cognition, 25*(5), 582–602. https://doi.org/10.1521/soco.2007.25.5.582.

Eggimann, S., Truffer, B., & Maurer, M. (2015). To connect or not to connect? Modelling the optimal degree of centralisation for wastewater infrastructures. *Water Research, 84*, 218–231. https://doi.org/10.1016/j.watres.2015.07.004.

Eisenführ, F., Weber, M., & Langer, T. (2010). *Rational decision making*. Berlin, Germany: Springer-Verlag.

Fanfar. (2021). FANFAR: Operational flood forecasting and alerts in West Africa. A transdisciplinary international project funded by the European Union. Retrieved from https://fanfar.eu and https://cordis.europa.eu/project/id/780118/results on 10 March 2021

Fetene, G. M., Olsen, S. B., & Bonnichsen, O. (2014). Disentangling the pure time effect from site and preference heterogeneity effects in benefit transfer: An empirical Investigation of transferability. *Environmental and Resource Economics, 59*(4), 583–611. https://doi.org/10.1007/s10640-013-9751-y.

Franco, L. A., & Hämäläinen, R. P. (2016). Behavioural operational research: Returning to the roots of the OR profession. *European Journal of Operational Research, 249*(3), 791–795.

Franco, L. A., Hämäläinen, R. P., Rouwette, E. A. J. A., & Leppänen, I. (2021). Taking stock of behavioural or: A review of behavioural studies with an intervention focus. *European Journal of Operational Research*. https://doi.org/10.1016/j.ejor.2020.11.031.

French, S. (2021). From soft to hard elicitation. *Journal of the Operational Research Society*, 1–17. https://doi.org/10.1080/01605682.2021.1907244.

French, S., Simpson, L., Atherton, E., Belton, V., Dawes, R., Edwards, W., et al., (1998). Problem formulation for multi-criteria decision analysis: Report of a workshop. *Journal of Multi-Criteria Decision Analysis, 7*(5), 242–262.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association, 32*(200), 675–701.

Gigerenzer, G., Hertwig, R., & Pachur, T. (2011). *Heuristics: The foundations of adaptive behavior*. Oxford University Press.

Gregory, R., Failing, L., Harstone, M., Long, G., McDaniels, T., & Ohlson, D. (2012). *Structured decision making: A practical guide to environmental management choices*. John Wiley & Sons.

Gregory, R., Lichtenstein, S., & Slovic, P. (1993). Valuing environmental resources: A constructive approach. *Journal of Risk and Uncertainty, 7*(2), 177–197.

Haag, F., Aubert, A. H., & Lienert, J. (2022). ValueDecisions, a web app to support decisions with conflicting objectives, multiple stakeholders, and uncertainty. *Environmental Modelling & Software, 150*, Article 105361. https://doi.org/10.1016/j.envsoft.2022.105361.

Haag, F., Lienert, J., Schuwirth, N., & Reichert, P. (2019a). Identifying non-additive multi-attribute value functions based on uncertain indifference statements. *Omega, 85*, 49–67. https://doi.org/10.1016/j.omega.2018.05.011.

Haag, F., Reichert, P., Maurer, M., & Lienert, J. (2019b). Integrating uncertainty of preferences and predictions in decision models: An application to regional wastewater planning. *Journal of Environmental Management, 252*, Article 109652.

Haag, F., Zürcher, S., & Lienert, J. (2019c). Enhancing the elicitation of diverse decision objectives for public planning. *European Journal of Operational Research, 279*(3), 912–928. https://doi.org/10.1016/j.ejor.2019.06.002.

Hämäläinen, R. P., Luoma, J., & Saarinen, E. (2013). On the importance of behavioral operational research: The case of understanding and communicating about dynamic systems. *European Journal of Operational Research, 228*(3), 623–634. https://doi.org/10.1016/j.ejor.2013.02.001.

Hayashi, K., Hondo, H., & Moriizumi, Y. (2016). Preference construction processes for renewable energies: Assessing the influence of sustainability information and decision support methods. *Sustainability, 8*(11), 1114.

Hoffmann, S., Feldmann, U., Bach, P. M., Binz, C., Farrelly, M., Frantzeskaki, N., et al., (2020). A research agenda for the future of urban water management: Exploring the potential of nongrid, small-grid, and hybrid solutions. *Environmental Science & Technology, 54*(9), 5312–5322. https://doi.org/10.1021/acs.est.9b05222.

Jacques, J., Grimonprez, Q., & Biernacki, C. (2014). Rankcluster: An R package for clustering multivariate partial rankings.

Janis, I. L. (1971). Groupthink. *Psychology Today, 5*(6), 43–46.

Katsikopoulos, K. V., Durbach, I. N., & Stewart, T. J. (2018). When should we use simple decision models? A synthesis of various research strands. *Omega, 81*, 17–25. https://doi.org/10.1016/j.omega.2017.09.005.

Keeney, R. L., & Raiffa, H. (1976). *Decisions with multiple objectives: Preferences and*. Cambridge & New York: Cambridge University Press.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika, 30*(1/2), 81–93.

Kendall, M. G., & Smith, B. B. (1939). The problem of m rankings. *The Annals of Mathematical Statistics, 10*(3), 275–287.

Kerr, N. L., & Tindale, R. S. (2004). Group performance and decision making. *Annual Review of Psychology, 55*(1), 623–655. https://doi.org/10.1146/annurev.psych.55.090902.142009.

Korhonen, P., & Wallenius, J. (1996). Behavioural issues in MCDM: Neglected research questions. *Journal of Multi-Criteria Decision Analysis, 5*(3), 178–182.

Larsen, T. A., Hoffmann, S., Lüthi, C., Truffer, B., & Maurer, M. (2016). Emerging solutions to the water challenges of an urbanizing world. *Science, 352*(6288), 928–933 (New York, N.Y.). https://doi.org/10.1126/science.aad8641.

Lichtenstein, S., & Slovic, P. (2006). *The construction of preference*. Cambridge University Press.

Liebe, U., Meyerhoff, J., & Hartje, V. (2012). Test–retest reliability of choice experiments in environmental valuation. *Environmental and Resource Economics, 53*(3), 389–407. https://doi.org/10.1007/s10640-012-9567-1.

Lienert, J., Andersson, J. C. M., Hofmann, D., Silva Pinto, F., & Kuller, M. (2022). The role of multi-criteria decision analysis in a transdisciplinary process: Co-developing a flood forecasting system in western Africa. *Hydrology and Earth System Sciences, 26*(11), 2899–2922. https://doi.org/10.5194/hess-26-2899-2022.

Lienert, J., Duygan, M., & Zheng, J. (2016). Preference stability over time with multiple elicitation methods to support wastewater infrastructure decision-making. *European Journal of Operational Research, 253*(3), 746–760.

Lienert, J., Schnetzer, F., & Ingold, K. (2013). Stakeholder analysis combined with social network analysis provides fine-grained insights into water infrastructure planning processes. *Journal of Environmental Management, 125*, 134–148. https://doi.org/10.1016/j.jenvman.2013.03.052.

Loomis, J. B. (1989). Test-retest reliability of the contingent valuation method: A comparison of general population and visitor responses. *American Journal of Agricultural Economics, 71*(1), 76–84. https://doi.org/10.2307/1241776.

McConnell, K. E., Strand, I. E., & Valdés, S. (1998). Testing temporal reliability and carry-over effect: The role of correlated responses in test-retest reliability studies. *Environmental and Resource Economics, 12*(3), 357–374. https://doi.org/10.1023/A:1008264922331.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological), 42*(2), 109–127.

Montibeller, G., & Von Winterfeldt, D. (2015). Cognitive and motivational biases in decision and risk analysis. *Risk Analysis, 35*(7), 1230–1251.

Montibeller, G., & von Winterfeldt, D. (2018). Individual and group biases in value and uncertainty judgments. *Elicitation: The science and art of structuring judgement* (pp. 377–392) L. C. Dias, A. Morton, & J. Quigley. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-65052-4_15.

Moors, G., Vriens, I., Gelissen, J. P. T. M., & Vermunt, J. K. (2016). Two of a kind. Similarities between ranking and rating data in measuring values. *Survey Research Methods, 10*(1). https://doi.org/10.18148/srm/2016.v10i1.6209.

Mørkbak, M. R., & Olsen, S. B. (2015). A within-sample investigation of test–retest reliability in choice experiment surveys with real economic incentives. *Australian Journal of Agricultural and Resource Economics, 59*(3), 375–392. https://doi.org/10.1111/1467-8489.12067.

Morton, A., & Fasolo, B. (2009). Behavioural decision theory for multi-criteria decision analysis: A guided tour. *Journal of the Operational Research Society, 60*(2), 268–275.

Nka, B. N., Oudin, L., Karambiri, H., Paturel, J. E., & Ribstein, P. (2015). Trends in floods in West Africa: Analysis based on 11 catchments in the region. *Hydrology and Earth System Sciences, 19*(11), 4707–4719. https://doi.org/10.5194/hess-19-4707-2015.

Norström, A. V., Cvitanovic, C., Löf, M. F., West, S., Wyborn, C., Balvanera, P., et al., (2020). Principles for knowledge co-production in sustainability research. *Nature Sustainability, 3*(3), 182–190. https://doi.org/10.1038/s41893-019-0448-2.

Ormerod, R. J. (2014). The mangle of OR practice: Towards more informative case studies of 'technical' projects. *Journal of the Operational Research Society, 65*(8), 1245–1260. https://doi.org/10.1057/jors.2013.78.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1992). Behavioral decision research: A constructive processing perspective. *Annual Review of Psychology, 43*(1), 87–131.

Petty, R. E., Wegener, D. T., & Fabrigar, L. R. (1997). Attitudes and attitude change. *Annual Review of Psychology, 48*(1), 609–647. https://doi.org/10.1146/annurev.psych.48.1.609.

R Core Team. (2020). R: A language and environment for statistical computing, Version Vienna, Austria. Retrieved from https://www.R-project.org/

Riabacke, M., Danielson, M., & Ekenberg, L. (2012). State-of-the-art prescriptive criteria weight elicitation. *Advances in Decision Sciences, 2012*.

Roberts, R., & Goodwin, P. (2002). Weight approximations in multi-attribute decision models. *Journal of Multi-Criteria Decision Analysis, 11*(6), 291–303.

Rowe, G., & Wright, G. (1999). The Delphi technique as a forecasting tool: Issues and analysis. *International Journal of Forecasting, 15*(4), 353–375. https://doi.org/10.1016/S0169-2070(99)00018-7.

Schaafsma, M., Brouwer, R., Liekens, I., & De Nocker, L. (2014). Temporal stability of preferences and willingness to pay for natural areas in choice experiments: A test–retest. *Resource and Energy Economics, 38*, 243–260. https://doi.org/10.1016/j.reseneeco.2014.09.001.

Slovic, P., Finucane, M. L., Peters, E., & MacGregor, D. G. (2007). The affect heuristic. *European Journal of Operational Research, 177*(3), 1333–1352.

Von Winterfeldt, D., & Edwards, W. (1993). Decision analysis and behavioral research.

Warren, C., McGraw, A. P., & Van Boven, L. (2011). Values and preferences: Defining preference construction. *Wiley Interdisciplinary Reviews: Cognitive Science, 2*(2), 193–205.

Whitehead, J. C., & Hoban, T. J. (1999). Testing for temporal reliability in contingent valuation with time for changes in factors affecting demand. *Land Economics, 75*(3), 453–465. https://doi.org/10.2307/3147190.

Zeileis, A., Cribari, N.F., .Gruen, B., Kosmidis, I., Simas, A.B., .Rocha, A.V. et al. (2020). Package 'betareg'. R package.

Zheng, J., & Lienert, J. (2018). Stakeholder interviews with two MAVT preference elicitation philosophies in a Swiss water infrastructure decision: Aggregation using SWING-weighting and disaggregation using UTAGMS. *European Journal of Operational Research, 267*(1), 273–287.