Annals of Operations Research
https://doi.org/10.1007/s10479-024-05941-6

ORIGINAL RESEARCH

An experimental design for comparing interactive methods based on their desirable properties

Bekir Afsar[1] · Johanna Silvennoinen[1] · Francisco Ruiz[2] · Ana B. Ruiz[2] · Giovanni Misitano[1] · Kaisa Miettinen[1]

Received: 17 April 2023 / Accepted: 9 March 2024

Abstract

In multiobjective optimization problems, Pareto optimal solutions representing different tradeoffs cannot be ordered without incorporating preference information of a decision maker (DM). In interactive methods, the DM takes an active part in the solution process and provides preference information iteratively. Between iterations, the DM can learn how achievable the preferences are, learn about the tradeoffs, and adjust the preferences. Different interactive methods have been proposed in the literature, but the question of how to select the best-suited method for a problem to be solved remains partly open. We propose an experimental design for evaluating interactive methods according to several desirable properties related to the cognitive load experienced by the DM, the method's ability to capture preferences and its responsiveness to changes in the preferences, the DM's satisfaction in the overall solution process, and their confidence in the final solution. In the questionnaire designed, we connect each questionnaire item to be asked with a relevant research question characterizing these desirable properties of interactive methods. We also conduct a between-subjects experiment to compare three interactive methods and report interesting findings. In particular, we find out that trade-off-free methods may be more suitable for exploring the whole set of Pareto

✉ Bekir Afsar
bekir.b.afsar@jyu.fi

Johanna Silvennoinen
johanna.silvennoinen@jyu.fi

Francisco Ruiz
rua@uma.es

Ana B. Ruiz
abruiz@uma.es

Giovanni Misitano
giovanni.a.misitano@jyu.fi

Kaisa Miettinen
kaisa.miettinen@jyu.fi

[1] University of Jyvaskyla, Faculty of Information Technology, P.O. Box 35 (Agora), FI-40014 University of Jyvaskyla, Finland

[2] Department of Applied Economics (Mathematics), Universidad de Málaga, C/ Ejido 6, 29071 Málaga, Spain

Published online: 17 April 2024

optimal solutions, while classification-based methods seem to work better for fine-tuning the preferences to find the final solution.

Keywords Multiple criteria optimization · Interactive methods · Performance comparison · Empirical experiments · Human decision makers

1 Introduction

Multiobjective optimization methods support a decision maker (DM) in finding the best balance among (typically conflicting) objective functions that must be optimized simultaneously. The DM's preference information is required to find the most preferred solution (MPS) among the mathematically incomparable Pareto optimal solutions that have different tradeoffs (Hwang & Masud, 1979; Miettinen, 1999; Steuer, 1986). Based on the DM's role in the solution process, multiobjective optimization methods can be classified into no-preference, a priori, a posteriori, and interactive ones, where the DM does not participate in the solution process or provides their preference information (preferences, for short) before, after or during the solution process, respectively (Hwang & Masud, 1979).

Interactive methods (Chankong & Haimes, 1983; Miettinen et al., 2008, 2016; Steuer, 1986), in which the DM takes part in the solution process iteratively, have proven useful because the DM can learn about the tradeoffs among the objective functions and about the feasibility of their preferences (Belton et al., 2008). Accordingly, they can adjust their preferences between iterations until they find the MPS. Interactive methods are computationally and cognitively efficient because only solutions of interest are generated, and only a few solutions per iteration are shown to the DM during the solution process. Therefore, many interactive methods are available in the literature. They differ, e.g., in the type of preference information the DM specifies, the type of information shown to the DM, how new solutions are generated, what the stopping criterion is, etc.

Choosing an appropriate interactive method for one's needs necessitates assessing and comparing their properties and performance. Since a DM plays an important role in interactive methods, the performance highly depends on human aspects. Experimental studies with human participants have been conducted in the literature to assess and compare interactive methods (see the survey (Afsar et al., 2021) and references therein). For example, the level of cognitive load experienced by the DM was assessed in Kok (1986) and the methods' ability to capture the DM's preferences in Buchanan (1994) and Narasimhan and Vickery (1988). Furthermore, the DM's satisfaction was assessed in Brockhoff (1985), Buchanan (1994), Buchanan and Daellenbach (1987), Korhonen and Wallenius (1989), Narasimhan and Vickery (1988) and Wallenius (1975). According to Afsar et al. (2023), most papers lack information on experimental details, such as the questionnaire used (i.e., exact questions asked) and data collected. Thus, they cannot be replicated to compare other methods.

Recently, an experimental design and a questionnaire were proposed in Afsar et al. (2023) to assess the DM's experienced level of cognitive load, the methods' ability to capture preferences, and the satisfaction of the DM in the solution process. The questionnaire aimed at measuring some desirable properties characterizing the performance of interactive methods, according to Afsar et al. (2021). A proof-of-concept experiment was also conducted in Afsar et al. (2023) to compare the reference point method (RPM) (Wierzbicki, 1980) and synchronous NIMBUS (NIMBUS) (Miettinen & Mäkelä, 2006) with a within-subjects design, where a small number of participants solved the same problem with both methods

(in different orders). This experiment demonstrated how the experimental setup worked, but no conclusions about the assessment of the methods compared could be derived, given that the results were not statistically significant.

Not all the desirable properties of interactive methods listed in Afsar et al. (2021) have been assessed before. In this paper, we design a questionnaire assessing multiple desirable properties. Its foundation is based on Afsar et al. (2023). We investigate the following aspects of interactive methods: cognitive load, capturing preferences, responsiveness to the changes in the DM's preferences, overall satisfaction, and confidence in the final solution. To avoid tiring participants, we have selected a between-subjects design, in which each participant solves the problem with only one method. This allows comparing more methods with more questionnaire items offering a deeper understanding of users' perceptions of applying different methods.

Besides conducting an experiment with the proposed design, one more contribution of this paper is reporting the insights gained. We compare three interactive methods: the E-NAUTILUS method (Ruiz et al., 2015), NIMBUS, and RPM. E-NAUTILUS is a trade-off-free method from the NAUTILUS family (Miettinen & Ruiz, 2016). In these methods, the DM starts from an inferior solution and gradually approaches Pareto optimal ones. This means that the DM gains in all objective functions simultaneously without trading off throughout the solution process. Including a tradeoff-free method in the experiment enables testing whether the proposed questionnaire can assess the above-mentioned aspects (e.g., a tradeoff-free method should place less cognitive load on the DM). On the other hand, NIMBUS uses the classification of the objective functions as the type of preference information. In each iteration, the DM examines the objective function values at the current solution and classifies each function into one of the five classes, indicating whether the function value (1) should improve, (2) should improve until a desired aspiration level is reached, (3) is currently acceptable, (4) may be impaired until some lower bound, or (5) can change freely. A classification is valid if at least one objective function should be improved and at least one is allowed to impair its current value. The DM provides aspiration levels and lower bounds for classes 2 and 4, respectively, and can specify the number of new Pareto optimal solutions to be generated for the next iteration. Finally, in each iteration of RPM, the DM provides preference information as a reference point consisting of aspiration levels. With this information, the method generates $k + 1$ Pareto optimal solutions, where k is the number of objective functions.

In the experiment conducted, we involve a high number of participants, which increases the reliability of the results. Having statistically significant results allows us to derive interesting conclusions about the behavior of the methods compared with respect to the desirable properties considered (which was not possible in Afsar et al. (2023)). In addition, we also develop a user interface (UI) for E-NAUTILUS similar to those implemented for NIMBUS and RPM in Afsar et al. (2023).

To summarize, the main contribution of this paper is two-fold. First, we design a questionnaire that can be used for experiments both to assess the performance of an individual interactive method and to compare different ones. Second, we share findings and insights from our experiment comparing interactive methods of different types.

The remainder of the paper is organized as follows. In Sect. 2, we outline general concepts of multiobjective optimization and briefly describe the considered aspects of interactive methods. We propose the extensive questionnaire in Sect. 3. We then focus on the experiment and its analysis and results in Sect. 4. In Sect. 5, we discuss and summarize our findings. Finally, we draw conclusions in Sect. 6.

2 Background

When a number of k (with $k \geq 2$) conflicting *objective functions* $f_i : S \to \mathbb{R}$ have to be optimized simultaneously over a feasible set $S \subset \mathbb{R}^n$ of solutions or decision vectors $\mathbf{x} = (x_1, \ldots, x_n)^T$, we have a *multiobjective optimization problem* (MOP) of the form[1]:

$$\begin{aligned} &\text{maximize} && \{f_1(\mathbf{x}), \ldots, f_k(\mathbf{x})\} \\ &\text{subject to} && \mathbf{x} \in S. \end{aligned} \tag{1}$$

We have *objective vectors* $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_k(\mathbf{x}))^T$ for $\mathbf{x} \in S$ and a *feasible objective region* Z, which is the image of S in the objective space $\mathbb{R}^k$ (i.e., $Z = \mathbf{f}(S)$). Usually, finding a single optimal solution where all objective functions can reach their individual optima is not possible because of the degree of conflict among the objective functions. Instead, several Pareto optimal solutions exist, at which no objective function can be improved without deteriorating at least one of the others. A solution $\mathbf{x} \in S$ is said to be *Pareto optimal* if there is no other $\bar{\mathbf{x}} \in S$ such that $f_i(\bar{\mathbf{x}}) \geq f_i(\mathbf{x})$ for all $i = 1, \ldots, k$, and $f_j(\bar{\mathbf{x}}) > f_j(\mathbf{x})$ for at least one index j. Its objective vector $\mathbf{f}(\mathbf{x})$ is called a *Pareto optimal objective vector*. All Pareto optimal solutions form a *Pareto optimal set* E, and the corresponding objective vectors form a *Pareto optimal front* $\mathbf{f}(E)$. The ranges of the objective function values in the Pareto optimal front are defined by the ideal and nadir points, denoted by $\mathbf{z}^\star = (z_1^\star, \ldots, z_k^\star)^T$ and $\mathbf{z}^{\text{nad}} = (z_1^{\text{nad}}, \ldots, z_k^{\text{nad}})^T$, respectively. For $i = 1, \ldots, k$, their components are defined as follows: $z_i^\star = \max_{\mathbf{x} \in S} f_i(\mathbf{x}) = \max_{\mathbf{x} \in E} f_i(\mathbf{x})$, and $z_i^{\text{nad}} = \min_{\mathbf{x} \in E} f_i(\mathbf{x})$.

Pareto optimal solutions are incomparable in a mathematical sense, and preference information from a DM is required to identify the MPS as the final solution. Different ways of expressing preferences can be used (Luque et al., 2011; Miettinen, 1999; Ruiz et al., 2012), such as e.g., selecting the most desired/undesired solution(s) among a set of alternatives, performing pairwise comparisons, giving a reference point formed by desirable objective function values (known as aspiration levels) or providing preferred ranges for the objective functions.

As stated in Miettinen et al. (2008), two phases can often be observed in interactive solutions processes, a learning phase and a decision phase, which are performed pursuing different purposes. In the learning phase, the DM learns about the problem, the available feasible solutions, and the feasibility of their own preferences. After exploring different solutions, the DM finds a *region of interest* (ROI) formed by Pareto optimal solutions that satisfy them the most. Then, a refined search within the ROI follows in the decision phase until finally finding the MPS.

A large variety of interactive methods have been developed, see e.g., Meignan et al. (2015), Miettinen (1999), Miettinen et al. (2016) and references therein. To find a suitable method for solving a problem, we need information about the properties and performance of different methods. Nevertheless, what "performance" means for an interactive method, i.e., how well it supports the DM in finding the MPS, is still an open question, given that different aspects must be considered. Desirable properties describing the performance of interactive methods have been proposed in Afsar et al. (2021). The authors also recognized the need of developing improved means for comparing interactive methods and provided general guidelines to conduct experiments with DMs. Since the quantitative assessment of interactive methods involving DMs is not trivial, further research is needed (Afsar et al., 2021; López-Ibáñez & Knowles, 2015).

[1] For simplicity, all objectives are assumed to be maximized. In case any of them must be minimized, it can be transformed into the maximization form by multiplying by -1.

Because of the central role of a DM in interactive methods, attention must be devoted to humans, and one aspect to be evaluated is the cognitive load set on the DM during the solution process. The cognitive load refers to the amount of working memory resources required and used (Sweller, 1988) with three types of cognitive load. Intrinsic cognitive load is the inherent level of difficulty and effort associated with a certain topic (Chandler & Sweller, 1991). The inherent levels of difficulty inducing cognitive load depend on an individual's capacities in specific problem-solving contexts. Extraneous cognitive load is caused by the ways information is presented (Chandler & Sweller, 1991), and germane cognitive load refers to the effort in processing and creating a mental knowledge structures (i.e., schema) of the topic (Sweller et al., 1998).

Thus, cognitive effort should not be demanding, and tiredness or confusion should be avoided during the solution process. To this aim, a DM must be provided with easy-to-understand information, shown using comprehensive visualizations to decrease the possibility of extraneous cognitive load emerging. Avoiding long waiting times and assuring that the MPS is found in a reasonable number of iterations should also be promoted. In addition, the way the preference information is elicited and the method's responsiveness (i.e., its ability to generate solutions reflecting the provided preferences even if they were changed drastically) influence the cognitive load.

The level of satisfaction of a DM when applying an interactive method in practice is also a determinant for evaluating its performance, given that the solution process usually stops when the DM is sufficiently satisfied (Afsar et al., 2021). Nevertheless, a DM may be satisfied with the final solution but may not be willing to interact with the method again if they found the interactive solution process e.g., too demanding or too difficult to be understood. Therefore, it is important to distinguish between the DM's satisfaction with the overall solution process and their satisfaction and confidence with the final solution. To ensure that the DM is confident enough with a solution before stopping, the method must promote learning about the tradeoffs during the solution process to allow the DM to get convinced of having reached a solution reflecting their preferences.

When experimenting with humans, we need to validate that the measurement's constructs measure the phenomenon being studied (Cook & Campbell, 1979). A validated measurement is a research instrument that has been tested to produce consistent results in terms of reliability and capture the issue intended to be measured, indicating the validity of the measurement. Many validated measurements have been developed to examine human perceptions, e.g., a validated measurement of the NASA Task Load Index (NASA TLX) has been developed for measuring cognitive load or level of experienced workload when interacting with technology (Hart & Staveland, 1988). NASA TLX has been created to evaluate six characteristics of cognitive load (mental, physical, and temporal demands, frustration, effort, and performance) in human-computer interaction. All of them are rated on a low to high scale, except performance, whose scale is from good to poor. To the best of our knowledge, validated measurements of cognitive load have not been developed for the specific characteristics of our experiment.

The after-scenario questionnaire (ASQ) (Lewis, 1995) is a validated measurement of a 3-item scale to measure user satisfaction with a computer system. The ASQ was developed within human-computer interaction for usability testing but can be applied to assess problem-solving with interactive multiobjective optimization methods due to its general nature. The measurement items are not restricted to specific contexts in humans interacting with technology.

The interactive methods used in our study have been implemented in the DESDEO framework (Misitano et al., 2021), a Python-based modular, open-source software framework for

interactive methods. As a part of this study, we have developed appropriate web-based UIs. The details about the multiobjective optimization problem that participants solved in the experiment with the methods can be seen in Section S1 of the supplementary material available at http://www.mit.jyu.fi/optgroup/extramaterial.html. In addition, brief descriptions of the interactive methods used in the experiment (E-NAUTILUS, NIMBUS, and RPM) can be found in Section S2 of this supplementary material.

3 Questionnaire design

In this section, we first list our research questions with our reasoning behind them. We then describe the proposed questionnaire and discuss connections to the desirable properties of interactive methods, research questions, and existing validated measurements.

3.1 Research questions

In this paper, we aim to assess important aspects of interactive methods such as the level of cognitive load experienced by the DM, the method's ability to capture preferences, the method's responsiveness to changes in the DM's preferences, the satisfaction of the DM in the overall solution process, and the DM's confidence in the final solution. Accordingly, we have selected some desirable properties of interactive methods (Afsar et al., 2021) and connected them to our research questions (RQs) as presented in Table 1.

Desirable properties related to the cognitive load are grouped into *RQ1: Cognitive load*, which covers how extensive the level of cognitive load experienced by the DM was in the whole solution process. *RQ2: Capturing preferences and responsiveness* examines the method's ability to capture a DM's preferences and the method's responsiveness. With *RQ3: Satisfaction and confidence*, we aim to investigate a DM's confidence in the final solution and the satisfaction of the DM with the overall solution process.

3.2 Questionnaire

In this section, we propose our questionnaire in Table 2, classified based on the timing of asking the question (for short, *timing* is used in Table 2). Apart from the questions to be answered once the solution process is over, some questions are to be answered after some specific iterations of the solution process. This includes both statements to be graded on a given scale and open-ended questions to be answered in writing. For the sake of brevity, we will refer to them as *items* henceforth. In Table 2, we list the desirable properties with RQs we have in Sect. 3.1 to show the corresponding item's purpose in the column 'purpose' (i.e., which desirable property is being assessed). For example, item RQ2-1 investigates RQ2 and its first desirable property (in Table 1). We use a 7-point Likert scale (strongly disagree (1)–strongly agree (7)) (Joshi et al., 2015; Likert, 1932) or a 5-point semantic differential scale (e.g., very low (1)–very high (5)), enabling us to perform quantitative analysis. Moreover, in some items, participants first grade on a scale and then explain the reasoning behind their grades.

The first four items in Table 2 are to be asked during the solution process, while the remaining ones are to be asked after the solution process. We present the items in the order to be used in experiments. In the first iteration, items 1 and 2 are asked after the participants have provided their preference information for the first time and before seeing the corresponding

Table 1 Our research questions connected to desirable properties from Afsar et al. (2021)

Research questions	Corresponding desirable properties
RQ1: Cognitive load: How extensive is the cognitive load of the whole solution process?	1) "The method sets as low cognitive burden on the DM as possible." 2) "The method allows the DM to fine-tune solutions in a reasonable number of iterations and/or reasonable waiting time."
RQ2: Capturing preferences and responsiveness: How well does the method capture and respond to the DM's preferences?	1) "The method captures the preferences of the DM." 2) "The DM feels being in control while interacting with the method." 3) "The method easily changes the area explored as a response to a change in the preference information given by the DM."
RQ3: Satisfaction and confidence: Is the DM satisfied with the overall solution process and confident with the final solution?	1) "The method allows the DM to learn about the conflict degree and tradeoffs among the objectives in each part of the Pareto optimal set explored." 2) "The method does not miss any Pareto optimal solution that is more preferred (with a given tolerance) for the DM than the one chosen." 3) "The method allows the DM to be fully convinced that (s)he has reached the best possible solution at the end of the solution process."

solution(s) generated by the method, while items 3 and 4 are asked after they have seen the solution(s). Similarly, in the fourth iteration, item 2 is asked after they have provided the preferences, but before seeing the solution(s), and items 3 and 4 once they have seen the solution(s) computed based on the preferences provided (we do not ask these questions at every iteration to avoid overloading the participants). Note that item 1 is only asked in the first iteration. In what follows, we elaborate on the items according to the RQs we have in Sect. 3.1.

In assessing the level of cognitive load experienced by the DM (RQ1), we get inspiration from the NASA-TLX questionnaire (Hart, 2006; Hart & Staveland, 1988) discussed in Sect. 2. Items 14 and 25 assess the experienced level of mental demand; items 22 and 24 a DM's mental effort, and item 23 the frustration level of the DM. Besides, we have item 13 to assess the level of performance of the DM in finding the final solution, similar to the NASA-TLX measuring one's performance in a given task. NASA-TLX has two more measurements (physical demand concerning participants' physical activity level and temporal demand concerning the time pressure placed on participants) that are inapplicable to our context—solving a multiobjective optimization problem does not require physical activity, and we do not set any time restrictions in our experiments.

As mentioned in Sect. 2, another key aspect describing the performance of an interactive method is the method's ability to reflect the preferences of the DM (RQ2). In particular, during the learning phase, the DM can provide preferences to explore different solutions. This means that the preferences may differ drastically from one iteration to the next. Thus, the method's ability to generate solutions reflecting the DM's preferences is crucial. Items 1 and 15 aim to assess whether the DM could articulate preferences well during the solution process. We have items 2, 16, 18, 19, and 20 to assess the method's ability to capture the preferences in terms of making it easy for the DM to provide preferences and having the

Table 2 Questionnaire items classified based on the research questions. Item 1 is designed to be asked after the first iteration, while items 2, 3, and 4 are to be asked twice: After the first and the fourth iteration. Items written in italics are from Afsar et al. (2023)

Timing	Questionnaire items	Answer type	Purpose
During the solution process	1) *The preference information was easy to provide.*	Likert scale	RQ2-1
	2) What do you wish to achieve by providing this preference information?	Open-ended	RQ2-1
	3) The solution(s) I obtained reflects my preference information well.	Likert scale	RQ2-3
	4) After this iteration, I know more about the problem.	Likert scale	RQ3-1
After the solution process	5) *Why did you stop iterating?*	Open-ended	RQ3-3
	6) *I think that the solution I found is the best one.*	Likert scale	RQ3-1
	7)*What degree of conflict do you think exists among each pair of objectives? a) Among f1 and f2 b) Among f1 and f3 c) Among f2 and f3*	Very low (1) Very high (5)	RQ3-1
	8) If you imagined a desired solution in the beginning, how similar is it when compared to the final solution you obtained? Please describe why?	Very dissimilar (1) Very similar (5) Open-ended	RQ3-3
	9) It was easy to explore solutions with different conflicting values of the objective functions.	Likert scale	RQ2-3
	10) I obtained a clear idea of the values that the objectives (indicators) can simultaneously achieve.	Likert scale	RQ3-1
	11) I obtained a clear idea of the possible choices available similar to the solutions I was interested in.	Likert scale	RQ3-1
	12) Did some solution(s) surprise you? Why?	Open-ended	RQ3-2
	13) I am satisfied with my performance in finding the final solution. Please describe why?	Likert scale open-ended	RQ1-1
	14) *A lot of mental activity was required (e.g., thinking, deciding, and remembering).*	Likert scale	RQ1-1
	15) *It was easy to learn to use this method.*	Likert scale	RQ2-2
	16) I was able to reflect my actual preferences when providing the information required by the method. Please describe why?	Likert scale Open-ended	RQ2-1
	17) In general, the method reacted to the preference information I provided.	Likert scale	RQ2-3
	18) I felt I was in control during the solution process.	Likert scale	RQ2-2
	19) I felt comfortable using this interactive method.	Likert scale	RQ2-2

Table 2 continued

Timing	Questionnaire items	Answer type	Purpose
	20) The method has all the necessary functionalities.	Likert scale	RQ2-2
	21) I was able to return to previous solutions whenever I needed in the solution process.	Likert scale	RQ2-2
	22) *I had to work hard to find the final solution.*	Likert scale	RQ1-1
	23) *I felt frustrated in the solution process (e.g., insecure, discouraged, irritated, stressed).*	Likert scale	RQ1-1
	24) *It took too many iterations to arrive to the acceptable solution.*	Likert scale	RQ1-2
	25) *I felt tired.*	Likert scale	RQ1-2
	26) Overall, I am satisfied with the ease of completing this task.	Likert scale	RQ3-3
	27) Overall, I am satisfied with the amount of time it took to complete this task.	Likert scale	RQ3-3
	28) Overall, I am satisfied with the support information (online help, messages, documentation) when completing this task.	Likert scale	RQ3-3
	29) I am satisfied with the solution I chose. Please describe why?	Likert scale Open-ended	RQ3-2

necessary functionalities so that the DM could feel in control during the solution process. Finally, items 3, 9, 17, and 21 assess the method's responsiveness as the ability to react to the DM's preferences.

In this paper, we consider the satisfaction in the overall solution process and the satisfaction (and confidence) in the final solution separately (RQ3). We first assess whether the DM has learned enough about tradeoffs among the objective functions in the problem considered. This is important since the DM cannot be confident in the final solution if they have not learned enough about the problem. Items 4, 6, 7, 10, and 11 evaluate whether the DM has gained insight into the problem (learned enough) or not. As mentioned, a DM typically stops the solution process when satisfied with the solution(s) found (Afsar et al., 2021). But overall satisfaction is also important, and they may stop for other reasons (e.g., being tired or not finding their preferred solution). We have items 5, 26, 27, and 28 to understand the overall satisfaction. Items 26, 27, and 28 come from ASQ (introduced in Sect. 2). Items 8, 12, and 29 assess whether the DM is satisfied and confident with the final solution.

In our questionnaire, we have developed new items and selected some proposed in Afsar et al. (2023) to investigate the aforementioned aspects of interactive methods. Besides the items listed in Table 2, we assess the participants' involvement as DMs with the questions "*The problem was easy to understand. Please describe why?*" and "*The problem was important for me to solve. Please describe why?*", as in Afsar et al. (2023). These questions are important to understand whether the participants take the experiment seriously, which improves the reliability of the experiments.

4 The experiment

4.1 UI design

We implemented the UIs of the three considered interactive methods following the same design principles as in Afsar et al. (2023) and utilizing the DESDEO framework (Misitano et al., 2021). The most notable difference from our previous work was the inclusion of a UI for E-NAUTILUS and the integration of the questionnaire items into the UIs. To illustrate the experimental setting, we give a brief description of the E-NAUTILUS UI in this section. More detailed descriptions of the UIs and their implementations are given in Section S3 of the supplementary material at http://www.mit.jyu.fi/optgroup/extramaterial.html.

The E-NAUTILUS UI is shown in Fig. 1. On the left of the figure, the UI for E-NAUTILUS is shown, where the participants can explore the points generated by the method and choose the one they prefer. After choosing a point and iterating, the participant is shown questionnaire items related to the preferences (Table 2, RQ2-1 and RQ2-2) as shown on the right of Fig. 1. The questionnaire items are positioned so that they do not block the view of the UI, but while they are shown, interacting with the UI is not possible before each item is answered. Other questionnaire items presented during the solution process are shown in a similar way. The questionnaire items showed after the solution process (Table 2) are shown in the same environment as well.

4.2 Participants and procedure

The participants (N = 164, 61% female, 39% male, age range 18–28, mean M = 19 years, standard deviation SD = 2.2) involved in this experiment were students from the Faculty

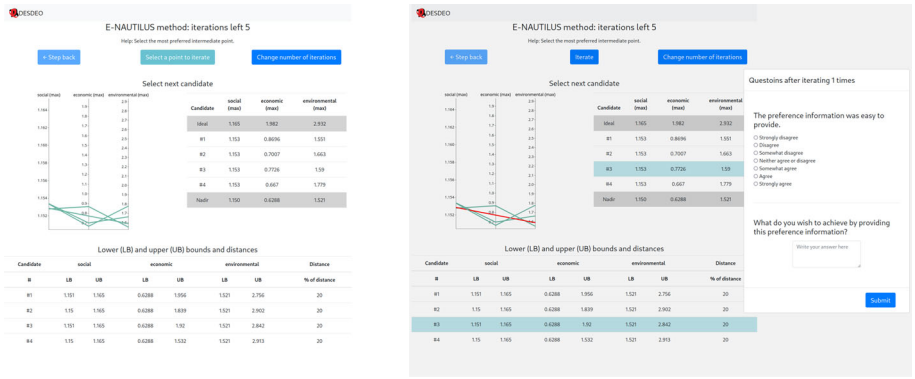

Fig. 1 Left: The UI of the E-NAUTILUS method. Right: Questionnaire items related to the given preferences as shown to the participant

of Economics and Business Studies of the University of Malaga. They all had very similar backgrounds in mathematics and multiobjective optimization. The participants were divided into three groups with one method assigned to each group, and the experiment was conducted in three separate sessions: one for E-NAUTILUS ($n = 64$), another for NIMBUS ($n = 44$), and another for RPM ($n = 56$). The numbers of participants differ because some students did not attend the experiment sessions they were assigned to.

We conducted a pilot study online (using the Zoom platform) with six participants (three of the co-authors and three collaborators) before the actual experiment. One co-author acted as an observer and another one as an experimenter, who started by presenting the informed consent, briefly describing the study and the procedure to be followed, and performing a live demonstration of the UIs (this took approximately 15 min). Then the experimenter sent to each participant (via chat) the web address of the UI, their credentials to log in, and the method to be used. The correct method was chosen for each participant based on the credentials. Each method was tested by two participants. After this pilot study, we could estimate how long the experiment would take. The general procedure was carried out as planned, and no modifications were needed.

Two weeks before the experiment, two co-authors described the main purpose and procedure of our study to the participants in each group in separate informative sessions. In these sessions, the general purpose of interactive multiobjective optimization methods was briefly presented, including a detailed description of the method to be used by each group. In addition, the multiobjective optimization problem to be solved was introduced, and a live demonstration of the UI of each method was made in the corresponding group. To let them think carefully about their preferences before the day of the experiment, we provided the students with supplementary documentation consisting of detailed information about the problem and the interactive method to be used.

At the beginning of the experiment, in each separate session, the participants were presented with informed consent. The experiment procedure was shortly reminded, and a tutorial video was shown to demonstrate the UI again. Besides, a 2-page printed summary of the problem and of the interactive method to be used was provided to let the participants recall the details during the experiment if needed. Next, they received their credentials printed on paper to log into the system's web address available on the course's virtual campus. As in the pilot study, the credentials ensured that they interacted with the appropriate method. Over-

all, the experiment took approximately 45 min in each session. It should be noted that the questionnaire was implemented in English. However, we provided the option to respond to open-ended questions in Spanish if the students found it more convenient to ensure that they could express their responses in a language that felt most comfortable for them. The supplementary documentation, the 2-page summaries of the problem and the interactive methods, and the video tutorials of the UIs can be seen in Sections S4, S5, and S6, respectively, of the supplementary material at http://www.mit.jyu.fi/optgroup/extramaterial.html.

4.3 Analysis and results

In what follows, we perform a quantitative and qualitative analysis of the participants' responses to the items of Sect. 3.2. As described in Sect. 4.1, they used radio buttons for the items answered in the Likert or semantic differential scales through the methods' UIs. They did not need to switch to another window to respond to the items, which allowed them to stay focused on the solution process and the questionnaire.

We applied the Kruskal–Wallis test (Kruskal & Wallis, 1952), a non-parametric test detecting statistically significant differences between the results of three or more independent groups. The test was appropriate for our needs because the items were measured on an ordinal scale (e.g., Likert scale and semantic differential) rather than a continuous scale, and we used the between-subjects design with three independent groups. For the p-values, the significance level was set at 0.05. In Table 3, we report the responses as mean scores (M) and standard deviations (SD) for each method, the latter in parentheses, along with the corresponding RQs. We also report the p-values of the Kruskal–Wallis test, with the statistically significant ones (less than 0.05) highlighted in bold in Table 3 (i.e., the corresponding item's results for the three methods differ significantly).

We analyzed the responses to the open-ended questions with a data-driven approach of qualitative content analysis (Weber, 1990). This included identifying semantic analysis units to create categories with iterative analysis. Data-driven qualitative content analysis requires an in-depth reading of the textual data with analysis iterations to create categories representing the data contents (Weber, 1990). Conducting data-driven qualitative content analysis can be time-consuming, but it is highly important and beneficial in gaining an in-depth understanding of participants' reasoning behind their numerical ratings. Next, we provide the quantitative and qualitative analysis for each RQ.

RQ1 – Cognitive load: In Table 3, the responses to the items related to RQ1 indicated the participants' experienced level of cognitive load for the whole solution process. The participants were significantly more satisfied with their performance in finding the final solution (item 13) using NIMBUS (M = 5.49; SD = 1.05) than with E-NAUTILUS (M = 5.21; SD = 1.15) or RPM (M = 4.79; SD = 1.45). Qualitative content analysis showed similar contents in describing the reasons behind the numerical ratings (item 13) for the question: *I am satisfied with my performance in finding the final solution. Please describe why?*. All the methods were evaluated in terms of whether a satisfactory solution was reached or not. Reasons for NIMBUS (N = 44) consisted of notions such as *"Yes, because it shows you the results according to the restrictions"*, for E-NAUTILUS (N = 64) e.g., *"I believe that I have been correctly evaluating the solutions as well as their possible consequences"*, and for RPM (N = 56) notions such as *"I am more or less in agreement with the solution that I have obtained"*. The only content-wise difference was that comments for RPM also included statements of how quick it was to obtain a solution (e.g., *"It was quite easy and quick"*).

Table 3 Responses as mean scores and standard deviations, the latter in parentheses. Item numbers in parentheses refer to items asked during the solution process ('it-1' means that the item was asked after the first iteration, and 'it-4' after the fourth iteration). Other items were asked after the solution process. Arrows indicate which direction is better (i.e., ↑ indicates that a higher value is better and ↓ means that a lower value is better). For items 7 and 8, the goodness cannot be indicated by higher or lower values but with semantic differentials (for item 7: very low (1)–very high (5); for item 8: very dissimilar (1)–very similar (5)). In the last column, *p*-values are listed, with statistically significant ones in bold

RQs	Items	E-NAUTILUS	RPM	NIMBUS	*p*-Value
RQ1-1	13 ↑	5.21 (1.15)	4.79 (1.45)	5.49 (1.05)	**0.0281**
	14 ↓	4.29 (1.51)	4.83 (1.19)	4.73 (1.60)	0.1153
	22 ↓	3.84 (1.36)	4.60 (1.46)	4.27 (1.55)	**0.0125**
	23 ↓	3.51 (1.64)	3.92 (1.99)	4.22 (1.52)	0.1198
RQ1-2	24 ↓	3.21 (1.46)	4.23 (1.71)	4.37 (1.83)	**0.0006**
	25 ↓	2.95 (1.85)	4.25 (1.91)	4.34 (1.84)	**0.0001**
RQ2-1	1 (it-1) ↑	5.52 (1.17)	5.13 (1.22)	5.25 (1.20)	0.1643
	16 ↑	5.29 (1.02)	4.96 (1.26)	5.32 (1.23)	0.2159
RQ2-2	15 ↑	5.87 (1.25)	5.13 (1.51)	5.37 (1.64)	**0.0163**
	18 ↑	4.84 (1.41)	4.94 (1.25)	4.73 (1.69)	0.9824
	19 ↑	5.38 (1.38)	5.19 (1.43)	5.46 (1.42)	0.5127
	20 ↑	5.13 (1.14)	4.94 (1.29)	5.46 (1.05)	0.1162
	21 ↑	5.35 (1.35)	4.11 (1.68)	5.02 (1.46)	**0.0001**
RQ2-3	3 (it-1) ↑	4.97 (1.44)	4.05 (1.76)	5.34 (1.22)	**0.0004**
	3 (it-4) ↑	5.05 (1.28)	4.03 (1.81)	5.29 (1.30)	**0.0084**
	9 ↑	5.25 (1.08)	4.40 (1.54)	4.80 (1.63)	**0.0110**
	17 ↑	5.13 (1.21)	4.89 (1.25)	5.46 (1.31)	**0.0267**
RQ3-1	4 (it-1) ↑	5.23 (1.24)	4.91 (1.53)	5.41 (1.00)	0.2861
	4 (it-4) ↑	5.23 (1.31)	4.74 (1.70)	5.21 (1.14)	0.4110
	6 ↑	4.33 (1.66)	4.28 (1.66)	5.05 (1.48)	**0.0378**
	7-a	3.02 (0.85)	3.09 (1.11)	2.98 (0.91)	0.6023
	7-b	2.95 (1.02)	3.00 (1.02)	3.22 (0.99)	0.3876
	7-c	4.00 (1.08)	4.15 (0.84)	3.90 (1.11)	0.6937
	10 ↑	5.51 (0.95)	5.06 (1.31)	5.20 (1.23)	0.1826
	11 ↑	5.52 (0.98)	5.08 (1.34)	5.46 (1.03)	0.1225
RQ3-2	29 ↑	5.14 (1.53)	4.81 (1.52)	5.68 (1.21)	**0.0068**
RQ3-3	8	3.27 (0.90)	3.15 (0.93)	3.15 (1.04)	0.7427
	26 ↑	5.48 (1.08)	5.00 (1.22)	5.44 (1.32)	**0.0306**
	27 ↑	5.37 (1.32)	4.83 (1.30)	5.05 (1.50)	0.0513
	28 ↑	5.44 (1.12)	5.17 (1.31)	5.20 (1.31)	0.5397

However, according to the statistical results above, this did not lead to a higher level of performance satisfaction in the RPM participants.

According to the responses to the mental activity (item 14) and the efforts in finding the final solution (item 22), RPM required more mental activity (M = 4.83; SD = 1.19) and effort (M = 4.60; SD = 1.46) than the other two methods. E-NAUTILUS required slightly less mental activity (M = 4.29; SD = 1.51) and significantly less effort (M = 3.84; SD =

1.36). Similarly, participants reported less frustration in the solution process (item 23) with E-NAUTILUS (M =3.51; SD = 1.64). However, even though RPM required more mental activity and effort, they were more frustrated with the NIMBUS solution process (M = 4.22; SD = 1.52). The participants felt that they needed more iterations with NIMBUS and RPM to arrive at an acceptable solution (item 24), and their tiredness level (item 25) was nearly the same. On the other hand, even though all E-NAUTILUS users reached the 4th iteration while roughly 45% of NIMBUS users and 36% of RPM users did not, they felt that the solution process with E-NAUTILUS required significantly fewer iterations (M = 3.21; SD = 1.46) and was less tiring (M = 2.95; SD = 1.85). Besides, the average time spent (in seconds) with E-NAUTILUS (M = 966.26; SD = 257.78) was higher than with NIMBUS (M = 844.11; SD = 380.76) and with RPM (M = 812.82; SD = 304.78). We can only report here the total time, but it is important in future experiments to record separately the time the DM spent interacting with the method, the computing time, and the time used in answering the questionnaire.

RQ2 – Capturing preferences and responsiveness: The ability of the methods to capture and respond to preferences was assessed using the RQ2 items in Table 3. Item 1 was only asked once, after the first iteration. The mean scores for all methods indicate that the participants easily provided preferences in all methods. They did, however, find E-NAUTILUS (M = 5.52; SD = 1.17) to be slightly easier than the others. Answers to the question *I was able to reflect my actual preferences when providing the information required by the method. Please describe why?* (item 16) were reasoned similarly for all the methods (e.g., E-NAUTILUS (N = 64): *"I have been able to choose according to my preferences at any moment"*, NIMBUS (N = 44): *"The programme has been able to interpret the data that I have entered"*, and RPM (N = 56) *"The method is good enough in order to represent user's preferences in these three aspects"*. All the methods also included notions of difficulties in reflecting preferences (e.g., E-NAUTILUS: *"It wasn't that easy to provide my preferences due to the conflicting objectives"*, NIMBUS: *"Because I wanted to increase the economic dimension, but the application did not increase it as I wanted it to"*, and RPM: *"Because the economic dimension affects too much the environmental one"*). One can say that these comments indicate a need to learn about trade-offs.

The participants found learning to use E-NAUTILUS significantly easier (item 15) (M = 5.87; SD = 1.25) than the other methods. They felt in control (item 18) and comfortable (item 19) during the solution process with all methods, and the methods provided all the necessary functionalities (item 20). However, when they wanted to return to previous solutions (item 21), E-NAUTILUS (M = 5.35; SD = 1.35) performed significantly better than the other two methods.

Written descriptions to the question *What do you wish to achieve by providing this preference information?* (item 2) were also analyzed with the qualitative content analysis (Weber, 1990) after the first and the fourth iteration. The answers of all the participants were similar. The focus was on either improving the preferred objective (e.g., for E-NAUTILUS and first iteration: *"My objective is to improve in first place the economy, and in second place the environment"* and same participant after the fourth iteration: *"I want to improve the three objectives, specifically my preferences are the economy and the environment"*), or on not to emphasize one objective but to seek a more balanced solution between the objectives (e.g., for E-NAUTILUS, *"I hope to find the best balanced solution"* and after the fourth iteration: *"I believe that a balance between the economic and environmental spheres represents an overall improvement for society"*). Overall, the rationale of what was wished to be achieved by providing preferences did not change between the first and the fourth iteration. The focus was more on fine-tuning the solution, either by increasing the value of one objective or by finding a more balanced solution.

The answers to the items related to the methods' responsiveness when the preferences changed significantly. Item 3 was asked during the solution process (after the first and fourth iterations); NIMBUS was the best at generating solutions that reflected participants' preferences well after the first iteration (M = 5.34; SD = 1.22), and this situation did not change after the fourth iteration (M = 5.29; SD = 1.30). Similarly, they felt that NIMBUS reacted best to their preferences (item 17) in general (M = 5.46; SD = 1.31). On the other hand, they found E-NAUTILUS to be significantly easier (M = 5.25; SD = 1.08) than the other methods in exploring solutions with different conflicting values (item 9).

RQ3 – Satisfaction and confidence: To determine which method was superior in terms of overall satisfaction and confidence in the final solution, we examined the responses given to the items for RQ3 in Table 3. During the solution process, knowing more about the problem (item 4) was measured twice. The participants' responses did not change significantly between the first and the fourth iterations, and they gained slightly more knowledge on the problem with E-NAUTILUS (M = 5.23; SD = 1.31) and NIMBUS (M = 5.21; SD = 1.14) than RPM (M = 4.74; SD = 1.70). They obtained a clear idea of the values that the objectives could simultaneously achieve (item 10), as well as possible choices available similar to the solutions they were interested in (item 11) with all methods. Moreover, with all the methods, they discovered that the second and the third objectives were in conflict with one another (item 7). When it comes to satisfaction and confidence in the final solution (item 29), NIMBUS outperformed the other two methods significantly. From the written descriptions to the question *I am satisfied with the solution I chose. Please describe why?* (item 29); the positive and negative statements were similar across all the methods. All of the comments pertained to reaching or not reaching a satisfactory solution. Positive example statements were given, such as *"Because it is the option I see as the best one of those obtained"*. For the participants who were not satisfied with the chosen solution, the reason was in not obtaining a satisfactory solution, e.g., *"I would have liked the environmental objective to be higher"*. Furthermore, they felt that the solution they found with NIMBUS was the best (item 6). Regarding overall satisfaction with the entire solution process (items 26, 27, and 28), E-NAUTILUS was slightly better than NIMBUS, and NIMBUS was slightly better than RPM.

For the question *If you imagined a desired solution in the beginning, how similar is it when compared to the final solution you obtained? Please describe why?* (item 8), 26 participants applying E-NAUTILUS (N = 64) stated that the final solution was similar to what they imagined (e.g., *"Because I have achieved a social level and an economic level very close to the ideal levels, without the environmental level being too much compromised since it would remain relatively as it is now"*). Differences between the final solution and the solution imagined in the beginning were described in terms of lower value than expected in the economic objective (n = 12/64), for example, with arguments such as *"My aim was to find a solution that would improve the economic indicator but without creating a great harm to the environment, but it was the objective that has suffered the most. The last variable in rank of importance was the social one, and this is the one that has experienced the highest improvement"*. The environment objective (n = 9/64) was imagined to have a higher value (e.g., *"The environmental dimension, whose range between the worst value and the best value is wider and therefore more difficult to approach to the most beneficial value, without considerably affecting the rest of the dimensions"*). The social objective (n = 6/64) was also considered to have influenced the final solution differently than what was imagined in the beginning. It was either considered as the most preferred objective aiming to be increased as much as possible or tried to be decreased (e.g., *"I wanted a low social factor, but it kept rising even if I tried keeping it low"*). Few participants (n = 6/64) stated that the final solution was

very different from the imagined one (e.g., *"Because I would have liked to achieve higher values for all three objectives but, given the problem and the conflict between the objectives, this was not possible"*), and 5 participants stated that they were not able to imagine a solution in the beginning.

Imagined solutions that corresponded to the solutions obtained were described by 17 participants applying NIMBUS ($N = 44$), with sentences such as *"My aims as described at the beginning are in line with what has been achieved"*. Differences compared to what was imagined were reported regarding higher values of the environment objective ($n = 5/44$), higher values of the economic objective ($n = 4/44$), and unexpected effects of the social objective to the other objectives ($n = 4/44$), e.g., *"I imagined that the social factor would remain between these values, which are very specific, and that the economic indicator would be quite dependent on the environmental one, and vice versa. That's why I'm looking for a point where I can improve the economic dimension by reducing the environmental dimension a little bit"*. Some participants ($n = 14/44$) also stated that the solution was very different from what they imagined (*"Because I thought the other two objectives would be in a better situation"*), or they could not imagine a solution before starting the solution process.

The same categories emerged from the analysis of responses given for RPM. From the participants ($N = 56$), only seven stated that the solution obtained was similar to the one they imagined. Differences between the imagined one and the obtained one were described according to unexpectedly low obtained values of the environment objective (n = 13/56), e.g., *"The environmental factor has to be strongly sacrificed in order to increase the other two"*, and of the economic objective (n = 12/56), e.g., *"Because I wanted to maximize the economic objective and it was impossible"*, and the social objective was considered as high ($n = 7/56$). From the participants applying RPM, 17 found the obtained solution very different (e.g., *"Since I wanted to maximize the economy, but I also wanted to have a high level of the environmental objective, but with high levels of economy, it's very difficult to have high levels of the environmental objective"*), or unimaginable (e.g., *"Because my imagined situation is not real"*).

Answers for the question *Did some solution(s) surprise you, why?* (item 12) were similar for all the methods. For E-NAUTILUS, a majority of the participants stated yes ($n = 44/64$), and the main reasons were due to unexpected results pertaining to the social objective ($n = 18/44$) (e.g., *"The social dimension reached very high values at the solutions from the beginning"*). Surprises regarding the environmental objective ($n = 9/44$) were described pertaining to its low value (e.g., *"Yes, the environmental one, by sacrificing this level too much to increase the other two"*), and, especially in relation to the economic objective (e.g., *"Yes, because I didn't know that if I wanted to maximize the economy, the environmental objective was so low"*). Unexpected changes in the economic objective were the smallest category ($n = 4/44$) (e.g., *"The economic function was very difficult to keep constant or to maximize"*). The second biggest category ($n = 13/44$) consisted of general statements depicting surprise of the interrelations of the objectives, such as *"Yes, the solutions that improve the economic and the environmental objectives, while the social one worsened"*. Participants who were not surprised ($n = 20/64$) expressed their reasons, for example with the following words *"No, in general, the solutions were within what I could expect"*.

For NIMBUS, less than half of the participants answered yes ($N = 19/44$), and 25 out of 44 participants said no. The reasons for encountering something unexpected pertained to lower values for the other objectives in increasing the social objective ($n = 4/19$) (e.g., *"Yes, in order to improve the social index, the economic one had to decrease a lot"*), to the low values of the environment objective ($n = 4/19$) (e.g., *"Improving the economy makes the environment worse"*), and due to the economic objective ($n = 4/19$) (e.g., *"Yes, for example, how the value*

of the economic index can change"). The second biggest category (n = 7/19) consisted of general statements *"Yes, because I didn't know that objectives could be so different from each other"*.

From the participants applying RPM (N = 56), 20 did not find surprises in the solutions, while the rest did (n = 36/56). Reasons for encountering something unexpected pertained to the high values in the social objective (n = 8/36) and were similar to the other methods, as well as lower values of the environment objective (n = 9/36), and especially in relation to the economic objective, with statements such as *"Yes, because the economic and environmental objectives are almost opposite in some cases"*. Only two participants reported surprises solely in the economic objective. The second biggest category (n = 17/36) of general statements included notions of surprises, such as *"Yes. Because they change a lot with a small variation"*, and *"Yes, because they have nothing in common with the ones I wanted to get"*.

Participants' involvement as DMs: The responses given to the problem-related questions were not significantly different. The problem was easily understood by all three groups (mean scores were over 5 on the Likert scale). However, the E-NAUTILUS users felt slightly more involved and found the problem important for them to solve (M = 5.27; SD = 1.18) than RPM (M = 4.75; SD = 1.58) and NIMBUS users (M = 4.83; SD = 1.45).

The written descriptions analyzed with qualitative content analysis revealed reasons behind the numerical scores given. Answers to the questions *The problem was easy to understand. Please describe why?* regarding: E-NAUTILUS were stated, such as *"Because it is a problem that is well embedded in today's society"* and *"Because the problem and its importance are very well explained"*. For the question *The problem was important for me to solve. Please describe why?*, responses when applying E-NAUTILUS were stated, such as *"Because it is something that affects everyone's life, and individually, with projects like this one, the situation can be better understood and help to some extent to solve it"*, for RPM e.g., *"Because it is about important aspects"*, and for NIMBUS e.g., *"It's a curious thing, it's important to understand how sustainability works"*.

Reasons for stopping iterating: The participants were asked to provide textual descriptions to the question: *Why did you stop iterating?* (item 5 in Table 2). The analysis resulted in two descriptive categories for all the methods: satisfactory solution found and misc/failure. The category of finding a satisfactory solution consists of descriptions of achieving the pursued goal as well as descriptions of reaching a good enough compromise between the objectives. The misc/failure category consists of notions of not being able to reach a satisfactory compromise and also utterances of not being able to apply the method correctly. In addition, for E-NAUTILUS, a third category was identified as a reason to stop iterating derived from the method: the number of iterations, in which reasons to stop iterating were defined according to the preset iteration rounds completed.

The majority of participants applying E-NAUTILUS (N = 64) stopped iterating due to reaching a satisfactory solution (n = 50/64). Justifications included statements such as *"I have found the most satisfactory balance according to my criteria"*. The number of iterations was stated as the stopping reason (n = 8/64), with a rationale such as *"Because I completed the number of iterations"*. The third category, misc/failure (n = 6/64), had reasons such as *"I couldn't get to the solution I was hoping for and I kept going in the same direction and found it difficult to straighten it out"*. For NIMBUS (N = 44), almost all the participants stopped iterating due to finding a satisfactory solution (n = 40/44). Reasons were stated, for example, *"I have managed to find the solution that most closely matches my preferences"*. Few participants stopped iterating because they did not reach a satisfactory solution (misc/failure (n = 4/44), with statements such as *"With different values, it gave me the same solution"*). Participants applying RPM (N = 56) mostly also stopped iterating due to finding a satisfactory

solution ($n = 40/56$). Reasons were described, such as *"Because I have found a solution that, more or less, maximizes the economy without reducing a lot the social and environmental values"*. Compared to E-NAUTILUS and NIMBUS, RPM gathered the most statements pertaining to misc/failure ($n = 16/56$). Descriptions for this stopping reason were described, for example, with the following words *"Because the program doesn't work well and I couldn't change the results"*.

5 Discussion

The experiment was designed as a between-subjects study enabling the comparison of three interactive methods with questionnaire items focusing on different aspects. When compared to a within-subjects design, where all the participants would have applied all the methods in a randomized order, it would not have been possible to ask as many questions due to an excessive workload impacting the results. A between-subjects design enabled designing the questionnaire in a way that included the assessment of many desirable properties of interactive methods. From the responses obtained, we can derive some interesting findings about the methods considered. In what follows, we discuss them in further detail.

First, the overall impression is that the participants were more confident in the final solution provided by NIMBUS, while they found E-NAUTILUS easier to use and less demanding. On the other hand, the RPM method did not appear to excel in any of the aspects considered. Interestingly, the method applied also influenced the perception of the participants about the problem to be solved. When asked to describe their involvement with the problem, the responses corresponding to RPM were more negative than for the other methods, even though the numerical results did not differ much between the methods. Going into further detail, the following observations can be made:

- The satisfaction with one's own performance was high for E-NAUTILUS and NIMBUS (the highest one for the latter) and a bit lower for RPM. Paradoxically, the highest frustration level also occurred with NIMBUS, while frustration was the lowest with E-NAUTILUS. A possible reason is that a tradeoff-free method requires lower mental activity, while NIMBUS makes the users more aware of the tradeoffs, especially in the ROI around the final solution. Besides, RPM seemed to require more mental activity (although the reported mental activity levels were high for all the three methods) and more hard work than the others, E-NAUTILUS being the best in these respects. Therefore, while the preference information required by RPM (reference points) is, in principle, simple to provide, it seems that the participants struggled to decide which information to provide to obtain satisfactory solutions.
- Surprisingly, despite the fact that all the participants iterated at least 4 iterations with E-NAUTILUS, while a significant percentage of them used fewer iterations with the other methods, and the overall time spent with E-NAUTILUS was longer, the impression of needing too many iterations and tiredness were much lower for E-NAUTILUS. These findings, once again, show that tradeoff-free iterations are cognitively less demanding and tiring. It must be said that the number of iterations is initially set in E-NAUTILUS (although it can be changed during the solution process). This may explain why the participants performed more iterations with this method, but their perceptions of time and tiredness are still interesting.
- The participants considered that E-NAUTILUS made it easier to explore solutions with different tradeoffs, but NIMBUS was the method that best reacted to their preference

information. In both cases, RPM was the worst one. Therefore, it seems that NIMBUS reflected preferences more accurately. The participants were able to correctly find the greatest conflict degree (between the economic f_2 and the environmental f_3 objective functions) with all three methods. In fact, many open-ended responses to the items *I was able to reflect my actual preferences when providing the information required by the method. Please describe why?* and *Did some solution(s) surprise you, why?* (see Sect. 4.3) prove this learning effect.

- E-NAUTILUS users reported having a clearer idea of the values that the objective functions could simultaneously achieve in the whole set of solutions, while they felt that NIMBUS performed better for identifying these values in the ROI, close to the final solution.
- A total of 43.2% of the participants applying NIMBUS re-started the solution process from the beginning, even more than once, while this percentage was 18% for RPM and just 9.4% for E-NAUTILUS. Maybe they did not understand properly what a feasible classification was in NIMBUS, although they reported that learning how to use this method was slightly harder than E-NAUTILUS but slightly easier than RPM. Better guidance in the NIMBUS UI could have been helpful.
- Despite previous responses about tiredness, mental activity, easiness to use, etc., NIMBUS users were more convinced than others that they had found the best possible solution. This may be explained by the fact that NIMBUS allows fine-tuning the final solution better than the other two methods. This interpretation is supported by the fact that there was greater satisfaction with the final solution obtained with NIMBUS than with others.

The above findings lead to an interesting conclusion: a tradeoff-free method like E-NAUTILUS seems appropriate for the beginning of the solution process, i.e., the learning phase, to allow the DM to explore the set of solutions without getting tired and stressed, and to determine one's ROI efficiently. On the other hand, a tradeoff-based method like NIMBUS, involving a classification, seems appropriate for the decision phase, where the DM can carry out a few more iterations to fine-tune the search and find one's MPS. This conclusion reinforces the appropriateness of building computational systems for interactive multiobjective optimization enabling the DM switching among different methods during the solution process (Heikkinen et al., 2023).

One possible limitation of this study is that the experiments were carried out with students, who are not expected to have a strong involvement with the problem. Although it would be hard to find such a large number of real DMs to carry out this experiment, we believe it would also be convenient to know the opinions of real DMs when using different methods. Another limitation is that the responses to the open-ended questions in Spanish were translated into English, which may have influenced the results of the qualitative analysis.

6 Conclusions

In this paper, we have proposed a questionnaire to assess and compare interactive methods corresponding to the following aspects: the DM's experienced level of cognitive load, the method's ability to capture and respond to preferences, the DM's overall satisfaction, and the DM's confidence in the final solution. We have carefully designed the questionnaire, i.e., the content, order, and timing of each item. In particular, apart from the items to be answered once the solution process is over, some items have been added in some specific iterations of the process to measure, e.g., the participants' learning about the tradeoffs during the solution

process. Furthermore, we deliberately chose the experimental design to be a between-subjects design, which allows asking many questions to assess the aforementioned aspects.

To demonstrate the applicability of the questionnaire, we conducted an experiment by comparing an interactive tradeoff-free method E-NAUTILUS and two more typical interactive methods, NIMBUS and RPM, that are based on dealing with Pareto optimal solutions throughout the solution process and focus more on tradeoffs. The methods compared were chosen with care to analyze the questionnaire in terms of the aspects considered. We were able to acquire useful outcomes. E-NAUTILUS, for example, was cognitively less demanding than the other methods, supporting the claim made for tradeoff-free methods. NIMBUS users, on the other hand, were more satisfied with the final solutions because they thought NIMBUS responded well to their preferences.

The proposed questionnaire, along with the experimental design and results, demonstrated its suitability for assessing and comparing interactive methods. We fully shared the questionnaire, which can be applied to future studies. Based on the results of this paper, we plan to conduct experiments to study the switch from one method to another during the solution process (e.g., one method at the beginning of the solution process to find the ROI (i.e., in the learning phase), and another method when the DM wants to fine-tune the solutions in the ROI (i.e., in the decision phase)).

Moreover, while the scope of this paper is focused on assessing the desirable properties of interactive methods, we acknowledge the value of exploring the specific solutions discovered during and at the end of the solution process. Therefore, as part of our future research studies, we plan to incorporate the analysis of solutions found, providing a more comprehensive understanding of the interactive methods and further enhancing the applicability of our findings.

Other future research directions include studying the role of the UI design solutions more in-depth affecting interaction in the solution process to improve the proposed UIs by decreasing possibilities of inducing extraneous cognitive load, as well as extending the questionnaire to further desirable properties discussed in the literature. It will also be interesting to develop validated measurements that are applicable for assessing the performance of interactive multiobjective optimization methods.

Supplementary information.

The supplementary material associated with this manuscript can be found at http://www.mit.jyu.fi/optgroup/extramaterial.html.

Acknowledgements This research is related to the thematic research area Decision Analytics utilizing Causal Models and Multiobjective Optimization (DEMO, jyu.fi/demo) at the University of Jyvaskyla, and was partly funded by the Academy of Finland (project 322221). This research was partly supported by the Spanish Ministry of Science (projects PID2019-104263RB-C42 and PID2020-115429GB-I00), the Regional Government of Andalucía (projects SEJ-532 and P18-RT-1566), and the University of Málaga (grant B1-2020-18).

Funding Open Access funding provided by University of Jyväskylä (JYU).

Data availability The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors declare that they have no Conflict of interest.

References

Afsar, B., Miettinen, K., & Ruiz, F. (2021). Assessing the performance of interactive multiobjective optimization methods: A survey. *ACM Computing Surveys, 54*(4), 1–27. Art. no. 85.

Afsar, B., Silvennoinen, J., Misitano, G., Ruiz, F., Ruiz, A. B., & Miettinen, K. (2023). Designing empirical experiments to compare interactive multiobjective optimization methods. *Journal of the Operational Research Society, 74*(11), 2327–2338.

Belton, V., Branke, J., Eskelinen, P., Greco, S., Molina, J., Ruiz, F., & Słowiński, R. (2008). Interactive multiobjective optimization from a learning perspective. In J. Branke, K. Deb, K. Miettinen, & R. Slowinski (Eds.), *Multiobjective Optimization: Interactive and Evolutionary Approaches* (pp. 405–434). Berlin: Springer.

Brockhoff, K. (1985). Experimental test of MCDM algorithms in a modular approach. *European Journal of Operational Research, 22*(2), 159–166.

Buchanan, J. T. (1994). An experimental evaluation of interactive MCDM methods and the decision making process. *Journal of the Operational Research Society, 45*(9), 1050–1059.

Buchanan, J. T., & Daellenbach, H. G. (1987). A comparative evaluation of interactive solution methods for multiple objective decision models. *European Journal of Operational Research, 29*(3), 353–359.

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction, 8*(4), 293–332.

Chankong, V., & Haimes, Y. Y. (1983). *Multiobjective Decision Making: Theory and Methodology*. North-Holland.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and Analysis Issues for Field Settings*. Houghton Mifflin Company.

Hart, S. G. (2006). NASA-task load index (NASA-TLX); 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 50*, 904–908.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (task load index): Results of empirical and theoretical research. *Advances in Psychology, 52*, 139–183.

Heikkinen, R., Sipila, J., Ojalehto, V., & Miettinen, K. (2023). Flexible data driven inventory management with interactive multiobjective lot size optimization. *International Journal of Logistics Systems and Management, 46*(2), 206–235.

Hwang, C.-L., & Masud, A. S. M. (1979). *Multiple Objective Decision Making - Methods and Applications: A State-of-the-Art Survey*. Springer.

Joshi, A., Kale, S., Chandel, S., & Pal, D. K. (2015). Likert scale: Explored and explained. *British Journal of Applied Science & Technology, 7*(4), 396–403.

Kok, M. (1986). The interface with decision makers and some experimental results in interactive multiple objective programming methods. *European Journal of Operational Research, 26*(1), 96–107.

Korhonen, P., & Wallenius, J. (1989). Observations regarding choice behaviour in interactive multiple criteria decision-making environments: An experimental investigation. In A. Lewandowski & I. Stanchev (Eds.), *Methodology and Software for Interactive Decision Support* (pp. 163–170). Berlin: Springer.

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association, 47*(260), 583–621.

Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction, 7*(1), 57–78.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 22*(140), 5–55.

López-Ibáñez, M., & Knowles, J. (2015). Machine decision makers as a laboratory for interactive EMO. In A. Gaspar-Cunha, C. Henggeler Antunes, & C. Coello Coello (Eds.) *Evolutionary Multi-criterion Optimization, 8th International Conference, Proceedings, Part II* (pp. 295–309). Springer.

Luque, M., Ruiz, F., & Miettinen, K. (2011). Global formulation for interactive multiobjective optimization. *OR Spectrum, 33*(1), 27–48.

Meignan, D., Knust, S., Frayret, J. M., Pesant, G., & Gaud, N. (2015). A review and taxonomy of interactive optimization methods in operations research. *ACM Transactions on Interactive Intelligent Systems, 5*(3), 17:1-17:43.

Miettinen, K. (1999). *Nonlinear Multiobjective Optimization*. Kluwer Academic Publishers.

Miettinen, K., Hakanen, J., & Podkopaev, D. (2016). Interactive nonlinear multiobjective optimization methods. In S. Greco, M. Ehrgott, & J. Figueira (Eds.), *Multiple Criteria Decision Analysis: State of the Art Surveys* (Vol. 2, pp. 931–980). Springer.

Miettinen, K., & Mäkelä, M. M. (2006). Synchronous approach in interactive multiobjective optimization. *European Journal of Operational Research, 170*(3), 909–922.

Miettinen, K., & Ruiz, F. (2016). NAUTILUS framework: Towards trade-off-free interaction in multiobjective optimization. *Journal of Business Economics, 86*(1), 5–21.

Miettinen, K., Ruiz, F., & Wierzbicki, A. P. (2008). Introduction to multiobjective optimization: Interactive approaches. In J. Branke, K. Deb, K. Miettinen, & R. Słowiński (Eds.), *Multiobjective Optimization: Interactive and Evolutionary Approaches* (pp. 27–57). Springer.

Misitano, G., Saini, B. S., Afsar, B., Shavazipour, B., & Miettinen, K. (2021). DESDEO: The modular and open source framework for interactive multiobjective optimization. *IEEE Access, 9*, 148277–148295.

Narasimhan, R., & Vickery, S. K. (1988). An experimental evaluation of articulation of preferences in multiple criterion decision-making (MCDM) methods. *Decision Sciences, 19*(4), 880–888.

Ruiz, A. B., Sindhya, K., Miettinen, K., Ruiz, F., & Luque, M. (2015). E-NAUTILUS: A decision support system for complex multiobjective optimization problems based on the NAUTILUS method. *European Journal of Operational Research, 246*, 218–231.

Ruiz, F., Luque, M., & Miettinen, K. (2012). Improving the computational efficiency in a global formulation (GLIDE) for interactive multiobjective optimization. *Annals of Operations Research, 197*(1), 47–70.

Steuer, R. E. (1986). *Multiple Criteria Optimization: Theory, Computation, and Applications*. Wiley.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257–285.

Sweller, J., Van Merrienboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*(3), 251–296.

Wallenius, J. (1975). Comparative evaluation of some interactive approaches to multicriterion optimization. *Management Science, 21*(12), 1387–1396.

Weber, R. P. (1990). *Basic Content Analysis*. Sage.

Wierzbicki, A. P. (1980). The use of reference objectives in multiobjective optimization. In G. Fandel & T. Gal (Eds.), *Multiple Criteria Decision Making, Theory and Applications* (pp. 468–486). Springer.