Sylvia Tulenvuo

DEVELOPING A SET OF SCALES TO ASSESS INTERACTIVE MULTIOBJECTIVE OPTIMISATION METHODS

JYVÄSKYLÄN YLIOPISTO
INFORMAATIOTEKNOLOGIAN TIEDEKUNTA
2024

ABSTRACT

Tulenvuo, Sylvia
Developing a set of scales to assess interactive multiobjective optimisation methods
Jyväskylä: University of Jyväskylä, 2024, 60 pp.
Cognitive science, master's thesis
Supervisor(s): Silvennoinen, Johanna; Miettinen, Kaisa; Kujala, Tuomo

Solving problems that involve considering multiple conflicting objective functions simultaneously is called multiobjective optimisation. Multiple mathematically equally good solutions can be found to these problems. These solutions are called Pareto optimal solutions. In order to choose one of these solutions as the final solution of the multiobjective optimisation problem considered, more information is required. This information is acquired from a decision-maker. The decision-maker is assumed to be an expert regarding the optimisation problem to be solved, and preference information provided by them is exploited to generate solutions that fit the decision maker's preferences. Multiobjective optimisation methods can be classified based on how the preference information is given. Methods where preference information is given progressively during the decision process are called interactive methods. Interactive methods repeat steps of the solution process until the decision-maker is satisfied and confident about the final solution. Interactive methods place a significant role on the decision-maker in solving the multiobjective optimisation problem. Despite the importance of the decision-maker, the literature lacks a validated measurement instrument to help develop and improve interactive methods to better match the needs and constraints of the decision-maker. The goal of this thesis was to examine whether research data from a previous study can be utilised to form a reliable scale or scales to assess interactive multiobjective optimisation methods. Principal component analysis was conducted to identify the components from the research data (N = 164). Three components were found: Cognitive load, Satisfaction and Decision-making support. Each component was calculated into a sum variable and their internal consistency was evaluated using Cronbach's alpha. Cronbach's alpha values were at an acceptable level. The correlations between the components indicate that they measure distinct constructs of interaction between the decision-maker and the interactive multiobjective optimisation methods and are therefore best utilised as individual scales. More research is needed in order to evaluate the validity of these scales.

Keywords: multiobjective optimisation, interactive multiobjective optimisation methods, decision-maker, scale development, principal component analysis

TIIVISTELMÄ

Tulenvuo, Sylvia
Vuorovaikutteisten monitavoiteoptimointimenetelmien arviointiin käytettävien mittareiden kehitys
Jyväskylä: Jyväskylän yliopisto, 2024, 60 s.
Kognitiotiede, pro gradu -tutkielma
Ohjaaja(t): Silvennoinen, Johanna; Miettinen, Kaisa; Kujala, Tuomo

Monitavoiteoptimointi on ongelmanratkaisua, jossa tulee yhtäaikaisesti ottaa huomioon useampi keskenään ristiriidassa oleva tavoitefunktio. Monitavoiteoptimoinnin ongelmiin on löydettävissä useita matemaattisesti keskenään yhtä hyviä ratkaisuja. Näitä ratkaisuja kutsutaan Pareto-optimaalisiksi ratkaisuiksi. Lisätietoa tarvitaan, jotta jokin näistä ratkaisuista voitaisiin valita ratkaistavan monitavoiteoptimointiongelman lopulliseksi ratkaisuksi. Tämä lisätieto saadaan päätöksentekijältä. Päätöksentekijän, jonka oletetaan olevan ratkaistavan monitavoiteoptimointiongelman asiantuntija, tarjoamaa preferenssitietoa hyödynnetään hänen preferensseihinsä sopivien vastausten luomiseksi. Monitavoiteoptimointimenetelmiä on useita erilaisia, ja ne voidaan luokitella sen mukaan, miten preferenssitietoa annetaan. Vuorovaikutteisissa monitavoiteoptimointimenetelmissä päätöksentekijä antaa preferenssitietoa pikkuhiljaa, ohjaten samalla päätöksentekoprosessia, kunnes hän on tyytyväinen lopulliseen ratkaisuun. Päätöksentekijällä on tärkeä rooli sopivan ratkaisun löytämisessä, mutta alalta puuttuu validoitu mittari, jonka avulla voitaisiin paremmin tutkia päätöksentekijää sekä päätöksentekoa vuorovaikutteisia monitavoiteoptimoinnin menetelmiä käytettäessä. Tässä tutkimuksessa oli tavoitteena selvittää, voiko aiemmasta tutkimusmateriaalista koostaa luotettavan mittarin tai mittareita, joilla voitaisiin arvioida vuorovaikutteisia monitavoiteoptimoinnin menetelmiä. Tutkimusdatan ($N = 164$) analysoinnissa hyödynnettiin pääkomponenttianalyysiä, jonka avulla tunnistettiin datasta löytyvät kolme komponenttia. Nämä kolme komponenttia nimettiin seuraavasti: Kognitiivinen kuormitus, Tyytyväisyys, ja Päätöksenteon tukeminen. Komponenteista laskettiin keskiarvosummamuuttujat, joiden sisäinen konsistenssi tarkastettiin Cronbachin alfan avulla. Komponenttien väliset korrelaatiot osoittavat, että ne mittaavat päätöksentekijän ja interaktiivisen monitavoiteoptimointimenetelmän välisen vuorovaikutuksen erillisiä osa-alueita, ja ovat täten parhaiten hyödynnettävissä yksittäisinä mittareina. Mittareiden validiteetin arviointi vaatii jatkotutkimusta.

Asiasanat: monitavoiteoptimointi, interaktiiviset monitavoiteoptimointimenetelmät, päätöksentekijä, mittarin kehitys, pääkomponenttianalyysi

FIGURES

TABLES

TABLE OF CONTENTS

1 INTRODUCTION

Real-world optimisation problems usually require consideration of different conflicting objectives. *Multiobjective optimisation* refers to problem-solving where multiple conflicting objective functions are considered simultaneously (Miettinen, 1999). Multiobjective optimisation has been applied in numerous fields, such as agriculture, banking, energy, farming, forestry, health services, insurance, military, and mining (White, 1990). This reflects the importance of multiobjective optimisation (Miettinen, 1999) and its applicability in today's society. *Interactive multiobjective optimisation methods* are iterative methods, where steps of the solution process are repeated until the final solution is found (Miettinen et al., 2008; Hwang & Masud, 1979). The performance of interactive multiobjective optimisation methods can be evaluated by how well they assist a decision-maker in finding the final solution (Afsar et al., 2021). The *decision-maker* (DM) is a person, who is the domain expert regarding the optimisation problem to be solved (Miettinen, 1999; Miettinen, 2008).

Despite the important role of the human DM in interactive methods, the literature lacks a profound understanding of the decision-making using interactive multiobjective optimisation methods. Comparisons between different interactive methods are infrequent (Afsar et al., 2021; 2023), user interfaces of interactive methods and the interaction between them and the DM is rarely mentioned in research (Afsar et al., 2021), and experiments conducted with human DM's are not reproducible because the detailed information of the study procedures is not available (Afsar et al., 2023). Additionally, no validated measurement instruments exist that could be used to assess interactive multiobjective optimisation methods. In order to advance the design, implementation and use of interactive multiobjective optimisation methods, more research is needed on the interaction between the DM and the interactive method user interface. A validated measurement scale is an easily applicable tool that can help designers and researchers to test their products quickly and efficiently, without developing a survey of their own.

This thesis has one primary objective, which is to examine if the research data from a previous study published in Afsar et al. (2024) can be utilised to form a reliable scale or scales to assess interactive multiobjective optimisation methods. The research question that guides the study in this thesis is the following:

> **RQ1**: Is the developed scale or a set of scales a reliable measure for assessing interactive multiobjective optimisation methods?

Because there are no validated scales in the field of *human-computer interaction* (HCI) or multiobjective optimisation that fit this purpose, the comparison between previously developed scales and the ones developed in this thesis cannot be done. Theories of human decision-making and judgment, and research on satisfaction and cognitive load conducted in the field of HCI are utilised to evaluate the validity of the developed scales.

The research data (N = 164) gathered by Afsar et al. (2024) was applied in this thesis to develop a set of measurement scales. A *principal component analysis* (PCA) was conducted, and parallel analysis was used to determine the number of components to retain. Three components were retrieved from the data. The components were named Cognitive load, Satisfaction, and Decision-making support, and calculated into sum variables. The internal coherence of each component is at an acceptable level. Correlations between the three components are relatively weak, hence they cannot be combined into a single scale. Each component can individually be used to reliably measure a distinct construct. Together they form a set of scales that measure important aspects of human decision-making in the context of utilising interactive multiobjective optimisation methods in decision-making.

The structure of this thesis is the following: the first two chapters, Chapters 2 and 3, present definitions of key concepts, including multiobjective optimisation, interactive multiobjective optimisation methods, theories of decision-making and judgment, and concepts of satisfaction and cognitive load in HCI. Chapter 4 describes the method, methodology, and the research material of the study reported here. Chapter 5 presents the results of the study and Chapter 6 includes the discussion of the results. Final conclusions are drawn in Chapter 7.

2 MULTIOBJECTIVE OPTIMISATION

This chapter covers the main concepts of multiobjective optimisation that are relevant for this thesis. At first, Chapter 2.1 introduces the concepts of Pareto optimal solutions and trade-offs, which are necessary for understanding the phenomenon of multiobjective optimisation. Chapter 2.2 addresses interactive multiobjective optimisation methods, and Chapter 2.3 presents the DESDEO software framework which incorporates methods regarding the study in this thesis.

2.1 Key concepts and terminology

Solving problems that involve considering multiple conflicting objective functions simultaneously is called multiobjective optimisation (Miettinen, 1999). A *multiobjective optimisation problem* of the form

$$\begin{aligned} &\text{minimise } \{ f_1(x), f_2(x), \ldots, f_k(x) \} \\ &\text{subject to } x \in S, \end{aligned}$$

where k ($k \geq 2$) *objective functions* $f_i : \mathbf{S} \rightarrow \mathbf{R}$ are simultaneously optimised (Miettinen, 1999), is considered. *Decision vectors* $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$ belong to the *feasible set* of solutions S, which is a subset of the *decision variable space* $\mathbf{R}^n$ (Miettinen, 1999). The vector of objective function values, denoted by $f(\mathbf{x}) = (f_1(x), f_2(x), \ldots, f_k(x))^T$, is called an *objective vector*. Objective functions are either minimised or maximised (Miettinen, 1999). If an objective function f_i is to be maximised, it is equivalent to minimise the function $- f_i$ (Miettinen, 1999). A solution that optimises each objective function does not exist, because the objective functions are conflicting (Miettinen, 1999). Therefore multiple, mathematically equally good solutions can be found (Miettinen, 2008). These solutions are known as *Pareto optimal solutions* that constitute a Pareto optimal set (Miettinen et al., 2008; Miettinen, 1999). A solution is Pareto optimal if improving the value of one objective function causes degeneration in at least one of the other objective functions (Miettinen, 1999; Miettinen et al., 2008). Therefore, trade-offs are necessary when compare Pareto optimal solutions to each other (Miettinen, 1999). Trade-offs are understood as follows: ”Trade-off reflects the ratio of change in the values of the objective functions concerning the increment of one objective function that occurs when the value of some other objective function decreases.” (Miettinen, 2008, s. 9).

The “best solution” cannot be absolutely defined amongst the Pareto optimal set (Rosenthal, 1985). Because these Pareto optimal solutions are mathematically incomparable in terms of which one is better, more information is needed to make a decision of selecting one solution candidate as the final solution of the problem considered. That information is expected to be obtained from a DM. In this thesis, the DM is defined following the Miettinen (1999; 2008) definitions: the DM is a person, who is assumed to know well the optimisation problem in question, and who can provide preference information regarding different solutions. *Preference information* is knowledge regarding the optimisation problem to be solved, that is not contained in the objective functions (Miettinen, 1999). Preference information is exploited to generate solutions that fit the DM’s preferences (Miettinen, 2008). The DM’s role is essential in finding the *most preferred solution* (MPS) (Miettinen et al., 2008). The DM does not need to know how multiobjective optimisation methods work, but they are expected to understand the problem (Miettinen, 2008).

2.2 Interactive methods of multiobjective optimisation

There are several ways to classify multiobjective optimisation methods (Miettinen, 1999). Hwang and Masud (1979) introduced a classification that is based on the type and timing of the given preference information. According to their classification, there are four kinds of methods:

- no preference information given
- preference information given a priori
- preference information given a posteriori
- preference information given interactively

No-preference methods are used when there is no preference information available (Miettinen, 1999; Miettinen, 2008). Therefore, the goal is to find a compromise solution “somewhere in the middle” of the Pareto optimal set (Miettinen, 2008; Wierzbicki, 1999). Components of *a neutral compromise solution* (Wierzbicki, 1999) are the average of the best and worst values of each objective function (Miettinen, 2008). In *a priori* methods, the preference information is given before the solution process (Hwang & Masud, 1979; Miettinen, 2008). The DM either accepts or rejects the obtained solution (Miettinen, 2008). In *a posteriori* methods the DM selects ones most preferred solution after the method has terminated and a set of

Pareto optimal solutions has been generated (Hwang & Masud, 1979). Methods where information is given progressively during the decision process are called interactive methods (Hwang & Masud, 1979). Interactive methods are iterative, meaning that the steps of the solution process are repeated until the DM is satisfied and confident about the final solution (Miettinen et al., 2008).

The goal of multiobjective optimisation is to support the DM's decision-making process and help them find the MPS (Buchanan, 1994; Xin et al., 2018). In addition, interactive methods also support the DM's learning process. One of their advantages is that they allow the DM to learn about the problem, the conflicting objective functions, and trade-offs between objective functions in different Pareto optimal solutions (Luque et al., 2011). Learning is an inherent part of interactive multiobjective optimisation (Belton et al., 2008). The DM does not know the set of Pareto optimal solutions beforehand: therefore, they have to learn what their preferred solution could be (Belton et al., 2008). The interaction process can even shape the DM's understanding so much, that the problem needs to be reformulated (Belton et al., 2008). According to Hwang and Masud (1979), interactive methods have a benefit of resulting in solutions that have a better prospect of being implemented. The DM participating in the solution process enhances the quality of the obtained solution. However, they require much more effort from the DM than no-preference and a priori methods (Hwang & Masud, 1979). But at the same time, interactive methods are cognitively less complex, since the DM can focus only on a small set of interesting solutions (Luque et al., 2011; Xin et al., 2018). It is also computationally faster to only generate solutions that interest the DM (Misitano et al., 2021).

2.3 Interactive methods of DESDEO software framework

The research data analysed in this thesis was originally collected for the study conducted by Afsar et al. (2024). They have proposed an experimental design to evaluate interactive multiobjective optimisation methods on their performance (Afsar et al., 2024). Three interactive methods were selected and compared: E-NAUTILUS, NIMBUS, and reference point method (RPM) (Afsar et al., 2024). These methods have been implemented in the DESDEO software framework (Misitano et al., 2021).

DESDEO is a modular, open-source software framework which has interactive multiobjective optimisation methods implemented in it (Misitano et al., 2021). It is implemented in the programming language Python, follows an object-oriented architecture design, and has several modules to implement interactive multiobjective optimisation methods that can also be used to modify existing ones (Misitano et al., 2021). DESDEO is unique in a sense that it is the only openly accessible framework in the field of multiobjective optimisation, that includes interactive methods (Misitano et al., 2021). The purpose of DESDEO is to provide a set of tools for researchers and practitioners to implement interactive methods and facilitate their development (Ojalehto & Miettinen, 2019).

DESDEO includes methods such as NIMBUS, NAUTILUS-family, and a reference point method (RPM). In the NIMBUS method, the DM classifies the objective function values of a current Pareto optimal solution in each iteration up to five preference classes (Miettinen & Mäkelä, 2006). After the DM has given their preferences, the method generates new Pareto optimal solutions and presents them to the DM. The DM decides the number of new solutions to be generated (Miettinen & Mäkelä, 2006). It is also possible to generate intermediate solutions between any two Pareto optimal solutions (Miettinen & Mäkelä, 2006). The search ends if the DM does not want to increase or decrease the value of any objective function (Miettinen & Mäkelä, 2006). In case of NIMBUS, the method starts with presenting the DM a Pareto optimal starting point (Miettinen & Mäkelä, 2006). The NAUTILUS family of methods is very different from that.

The NAUTILUS family contains interactive tradeoff-free methods (Miettinen & Ruiz, 2016). The idea of NAUTILUS methods is to support the DM's so-called "free search for the most preferred solution" (Miettinen et al., 2010). The solution process begins from the worst possible solution (Miettinen & Ruiz, 2016). From there, the DM iteratively gets closer to the set of Pareto optimal solutions and is able to improve all the objective function values simultaneously in each iteration (Miettinen & Ruiz, 2016). At the end, the DM will reach a Pareto optimal solution (Miettinen & Ruiz, 2016). The DM's motivation to iterate until the MPS is found may stay stronger since they do not have to sacrifice some objectives over the others but can keep improving the solutions towards more satisfactory at each iteration (Miettinen et al., 2010). The NAUTILUS method has been developed to diminish the effects of anchoring in human decision-making (Miettinen et al., 2010; Miettinen & Ruiz, 2016). Instead of starting with a Pareto optimal solution, it is more beneficial for the solution process to begin with an inferior solution, because in that case the DM always has a chance for a better solution (Miettinen et al., 2010). One of the methods in the NAUTILUS method family is E-NAUTILUS (Ruiz et al., 2015). It has been developed especially to solve computationally expensive problems (Miettinen & Ruiz, 2016). In E-NAUTILUS, first a set of Pareto optimal solutions is computed by using any a posteriori method e.g. some evolutionary multiobjective optimisation algorithms (Ruiz et al., 2015). After that, the DM chooses the next iteration point from a set of generated intermediate points which are all at a certain distance from the worst possible objective function values (Ruiz et al., 2015). After each iteration the intermediate points move closer to the previously generated set of Pareto optimal solutions until the DM finally reaches the most preferred one amongst them (Ruiz et al., 2015).

The third method used by Afsar et al. (2024) is the reference point method. In the reference point method, the DM is asked to define desired aspiration levels for all objective functions (Wierzbicki, 1982). Aspiration levels form a point that is called a reference point (Wierzbicki, 1980). The DM can change the aspiration levels in the solution process (Wierzbicki, 1980). The method generates Pareto optimal solutions based on reference points by using an achievement scalarizing function (Wierzbicki, 1980; Miettinen, 1999). Pareto optimal solutions generated and presented to the DM depend on the location of the reference point. The closer the reference point is to the Pareto optimal set, the more detailed description of the neighbourhood solutions is offered (Wierzbicki, 1980; Miettinen, 1999). According to Wierzbicki (1999; 1982), the reference point method is psychologically appealing because it is supported by Simon's concept of satisficing decisions, which is addressed later in this thesis in Section 3.1. Reference point -based methods are rather popular because of their straightforward nature (Luque et al., 2009). The concept of reference points is intuitive and the method easy to use (Miettinen, 1999).

3 DECISION-MAKING

This chapter covers the prominent theories of decision-making and judgment regarding decision-making with interactive multiobjective optimisation methods. The chapter also introduces factors that are present when decision-makers interact with decision-making systems, and the desirable properties of interactive multiobjective optimisation methods.

Decision-making is present in human personal and professional life in many ways. It is studied in different disciplines, including psychology, cognitive science, and economics. For many decades, research in judgment and decision-making focused on examining behaviour that violated the assumptions of rational choice theory (Mellers et al., 1998). Rational choice theory has greatly influenced the field of economics and all the behavioural and social sciences (Gilovich & Griffin, 2002; Simon, 1956). It has had a significant impact on research of human judgment and decision-making (Gilovich & Griffin, 2002). The theory assumes that rational people choose the option that maximises both the probability and utility of each possible outcome (Gilovich & Griffin, 2002). According to the rational choice theory, people perform probability and utility calculations all the time and they are good at it (Gilovich & Griffin, 2002).

Expected utility family of theories are the most historically dominant models of rational choice (Hastie, 2001). After the World War 2, the expected utility theory became the most influential theory of decision-making (Schoemaker, 1982), dominating the analysis of decision-making under risk and uncertainty (Kahneman & Tversky, 1979; Tversky, 1975). The theory of expected utility became generally accepted as a normative model of rational choice and was applied as a descriptive model of economic behaviour (Keeney & Raiffa, 1976 and Friedman & Savage, 1948, cited in Kahneman & Tversky, 1979), as well as a normative theory to aid in decision-making (Tversky, 1975).

Savage (1954) developed the theory of expected utility by adding the concept of subjective probability (Tversky, 1975). In contrast to the expected utility theory, "the probabilities are the decision-maker's personal or *subjective* probabilities for uncertain outcomes" (Frisch & Clemen, 1994; emphasis added). The theory of *subjective expected utility* (SEU) assumes that people make choices to maximise their subjective utility at each moment (Simon, 1990; Jones, 1999). When people make decisions, they choose from the predetermined set of options that are available to them (Savage, 1954, cited in Simon, 1990). The probability of each outcome is evaluated by the person's own subjective probabilities (Savage, 1954, cited in Simon, 1990). Even though the SEU theory has been the norm to explain human rational decision behaviour in past decades, it has also faced extensive criticism (Frisch & Clemen, 1994).

In all models of rational choice theory, it is assumed that rational people want to follow their principles and that they actually do so (Kahneman & Tversky, 1979). Theories assume that the DM has a complete knowledge of the consequences that will follow on each of the alternatives, or at least the DM is capable to compute them (Simon, 1979). According to these theories, the DM is able to compare even a diverse set of consequences against each other, and the comparisons are carried out by using the same measure of utility (Simon, 1979). It is assumed that the reason why the DM prefers some options over others is based on the fixed outcomes of these options (Jones, 1999). The DM knows what those outcomes are, and they maximise their utility by choosing an option with the highest benefit-to-cost ratio (Jones, 1999).

Despite having a strong influence, the models of rational choice theory and its assumptions have faced a lot of criticism. In the following chapter, theories that contradict or criticise the assumptions of rational choice theory are introduced. If the critique has been targeted at a specific variant of rational choice theory, such as expected utility theory or SEU, the specific variant is mentioned by its name. In other cases, the critique is assumed to be targeted at the general idea and assumptions behind the models of rational choice theory.

3.1 Theories of human decision-making and judgment

Research on bounded rationality began when Herbert Simon noticed that the assumptions of utility maximisation did not apply in real-world situations (Gigerenzer, 2020). Gigerenzer (2020) and Gigerenzer and Gaissmaier (2011) refer to Simon's writings from 1989 (p. 377):

> Now I had a new research problem: How do human being reason when the conditions for rationality postulated by the model of neoclassical economics are not met – for example, when no one can define the appropriate utility function, or suggest how the contribution of expenditures to utility is to be measured?

Bounded rationality originated in Simon's *Administrative Behavior* (1947), and it developed to be an alternative to the rational choice theory (Jones, 1999). Today, there exist several models of bounded rationality (Viale, 2020). Bounded rationality is "the most important idea - - that political science has ever exported" (Jones, 1999). It has influenced the fields of economics, psychology and management (Viale, 2020).

Simon's (1990) critique is especially targeted at the SEU theory. In his opinion, the SEU theory was a considerable improvement in rational choice theory, but similarly, it fails to demonstrate actual human decision-making behaviour because the assumption of perfect maximisation is still present in it, despite taking into account the subjective probabilities of a DM (Simon, 1979; Simon, 1990). The theory should rather be seen as a mean to predict choice (Simon, 1990), since an ordinary human mind cannot estimate the probabilities of even a small set of options when making a decision (Jones, 1999). The SEU theory still has an impact on Simon's work: the theories of bounded rationality can be derived from the SEU theory, if its assumptions are modified (Simon, 1990). So, instead of having a predetermined set of alternatives from which the DM chooses, the bounded rationality assumes a process for generating those alternatives (Simon, 1990). It acknowledges the cognitive capabilities and limitations of the human mind, while considering the environmental factors (Simon, 1990).

Simon did not reject the idea of human rationality altogether. The idea of "approximate rationality" characterises an organism with limited computational resources and information (Simon, 1956), in oppose to the omniscient DM in models of rational choice theory. The opposite of maximising in Simon's work is satisficing. As Viale (2020) describes it, the human rationality is "bounded, heuristic, and satisficing". Simon developed the idea of satisficing organism, that would satisfice, not optimise (Simon, 1955; 1956). Rational decision-making is usually defined as maximising utility (Simon, 1997, s. 195). In that context, the DM, who chooses the best alternative available, *optimises* (Simon, 1997, s. 295). In real life, it is not feasible to compute genuine optima (Simon, 1997, s. 295). Satisficing decision-making means that the DM "chooses an alternative that meets or exceeds specified criteria" (Simon, 1997, s. 295). The alternative is not necessarily unique or the best choice (Simon, 1997, s. 295).

The interactive RPM method discussed in Chapter 2 is based on Simon's idea of satisficing. As mentioned, in multiobjective optimisation, there is no unique solution that could be said to be "the best". The DM is obligated to make compromises. Satisficing makes sense especially in situations where optimizing (finding the genuine maximum or minimum) is not possible and when it is computationally too costly to do that (Simon, 1997, s. 295). In those cases, the DM searches for a satisfactory choice (Simon, 1997, s. 295). Simon describes the process of setting the criteria level for "satisfactory" as follows:

> Psychology proposes the mechanism of aspiration levels: if it turns out to be very easy to find alternatives that meet the criteria, the standards are gradually raised; if search continues for a long while without finding satisfactory alternatives, the standards are gradually lowered. (Simon, 1997, s. 296).

How easy or difficult it is to find the satisfactory alternative informs the DM of the applicability of their criteria. This modification of criteria eventually leads the DM to *converge* towards a set of feasible criteria (Simon, 1997, s. 296). Following the aspiration level mechanism is computationally simpler than optimizing (Simon, 1997, s. 296). The same goes for the RPM method.

Where theories of bounded rationality acknowledge that humans are capable of making rational decisions even if their knowledge is limited, heuristics and biases approach regards the human decision-making process as being fundamentally flawed. Behind the widely spread heuristics and biases approach is the research made by Kahneman and Tversky in the 1960's and 1970's (Gilovich & Griffin, 2002; Gigerenzer, 1991). A *cognitive bias* is "a systematic discrepancy between the (average) judgment of a person or a group and a true value or norm" (Gigerenzer, 2018). According to Tversky and Kahneman (1974), cognitive biases "stem from the reliance on judgmental heuristics" and they "lead to systematic and predictable errors". These biases persist even when the subjects are rewarded for correct answers or encouraged to be accurate (Tversky & Kahneman, 1974). In Tversky and Kahneman's (1974) opinion, this proves that biases cannot be eliminated by altering motivational effects. Cognitive biases can be therefore seen as unavoidable features of human decision-making and judgment.

Instead of following the statistical rules of reasoning, people evaluate the likelihood of events based on subjective assessments and intuitive heuristics (Tversky & Kahneman, 1974; 1977). Even though heuristics often work well enough for people to keep using them, they also lead to systematic errors in reasoning (Tversky & Kahneman, 1974; 1973). People are not capable of computing the probabilities of (uncertain) events (Tversky & Kahneman, 1977), which leads them to exchange the laws of chance for heuristics (Kahneman & Tversky, 1972). Tversky and Kahneman (Tversky & Kahneman, 1974; Kahneman & Tversky, 1972) came to a conclusion that breaking the rules of probability and statistic inference is systematic, predictable, and hard to eliminate. Based on their experiments, deeply rooted intuitions about uncertainty do not change even when people are given statistical training (Kahneman & Tversky, 1973). The deviation from the principles of probability theory when evaluating the likelihood of uncertain events is, by their words, "not surprising", because "the laws of chance are neither intuitively apparent, nor easy to apply" (Kahneman & Tversky, 1972).

Tversky and Kahneman demonstrated that there are three common heuristics – representativeness, availability, and anchoring and adjustment – that people apply when making predictions and judgments about uncertainty (Tversky & Kahneman, 1974; 1972; 1973; Kahneman & Tversky, 1972; 1973). When people predict what is going to happen in the future or what the outcome of certain action is, they evaluate the outcomes based on how well they represent the essential characteristics of the given evidence (Kahneman & Tversky, 1972; 1973). A simple example of the *representativeness* heuristic given by Kahneman and Tversky (1972) is evaluating the probability that a 12-year-old boy will become a scientist. The evaluation could be affected by how well the role of a scientist is representative of peoples' image of the boy (Kahneman & Tversky, 1972). Representative outcomes can sometimes be more likely than others, but because people ignore the other factors that affect the likelihood of certain events but not their representativeness, relying on this heuristic can lead to errors of judgment (Kahneman & Tversky, 1973).

The similar effect happens with *availability* heuristic, where people assess the frequence of a class or a probability of an event by how easily a certain occurrence or an instance can be recalled (Tversky & Kahneman, 1973; 1974). It is easier to remember frequent events than infrequent ones, naturally (Tversky & Kahneman, 1973; 1974). But when people make inferences using an availability heuristic, they also consider irrelevant factors that do not actually affect the likelihood or frequency of a certain event or a class (Tversky & Kahneman, 1973). It is a similar effect than when using a representativeness heuristic: people do not recognize the effect of other factors in their subjective assessments (Tversky & Kahneman, 1973).

In the context of multiobjective optimisation, *anchoring and adjustment* plays a considerable role. Buchanan and Corner's (1997) experimental evidence suggests that the anchoring bias affects the use of interactive multiobjective optimisation methods. The aforementioned NAUTILUS method is designed to take into account human decision-making biases (Miettinen & Ruiz, 2016). Anchoring describes a phenomenon where people estimate the final answer to be close to the initial starting point value (Tversky & Kahneman, 1974). The bias occurs even when the initial value is randomly generated and has nothing to do with the actual correct answer (Tversky & Kahneman, 1974). People fail to adjust their estimates to better match the correct answer, but instead anchor on the starting point value (Tversky & Kahneman, 1974). The NAUTILUS method avoids the anchoring effect by starting with an inferior solution so that the DM would not focus too much on attaining a solution that resembles the Pareto optimal solution(s) introduced first (Miettinen & Ruiz, 2016).

Tversky and Kahneman seem to have collected a lot of evidence to support the inevitability of the use of heuristics and them leading to cognitive biases. However, there are other viewpoints. In Gigerenzer's opinion (2018), Tversky and Kahneman's findings inspired to interpret a "variety of deviations from rational choice theory as systematic flaws in the human mind rather than in the theory". In behavioural economics, cognitive biases are held as truth (Gigerenzer, 2018). Regardless of other knowledge (e.g., in experimental psychology), the field of behavioural economics adopted a world view where people systematically make errors of judgment (Gigerenzer, 2018). This is not only true with economics: Gigerenzer (1991) notes that the influence of research on cognitive biases can be seen in social psychology, law, management science, medical diagnosis, and many other fields. Heuristics became associated with errors, but in certain decision-making situations they can be more accurate than "rational" strategies (Gigerenzer & Gaissmaier, 2011). Heuristics are not good or bad, rational or irrational: they should be interpreted in their context of use (Gigerenzer & Gaissmaier, 2011).

The prominence of cognitive biases has also been questioned by Gigerenzer. Gigerenzer (2018) defines the term *bias bias* as "the tendency to see systematic biases in behaviour even when there is only unsystematic error or no verifiable error at all". He reviews a good amount of literature to show that the alleged biases do not demonstrate themselves when the experiments are reformulated and modified to consider the statistical principles of the given context (Gigerenzer, 2018). Gigerenzer (2018) claims that especially in behavioural economics, the bias bias distorts the view of human decision-making as stubbornly erroneous. The findings behind the bias bias -theory have been established in Gigerenzer's earlier work from 1991, where he examined the research made of overconfidence bias, conjunction fallacy, and base-rate neglect, which have all been labelled errors in probabilistic reasoning. The results of the literature review revealed that the cognitive biases disappear when the assignment are reformulated to take into account the conceptual differences in probability (Gigerenzer, 1991). Gigerenzer (1991) also criticises the use of the phrase "the normative theory of prediction" in the heuristics and biases approach. In his opinion, it seems to mean that the subjects are assumed to mechanically apply a (Bayes's theorem) formula to solve the given problem (such as engineer-lawyer problem) (Gigerenzer, 1991). There are, however, other theories of probability, like frequentist, which in his words, is the most popular theory of probability (Gigerenzer, 1991). The evidence on all three examples seems to suggest that the human mind is actually a frequentist, that "distinguishes between single events and frequencies in the long run – just as probabilistics and statisticians do" (Gigerenzer, 1991). People are also sensitive to the difference between random vs. selected (non-random) samples (Gigerenzer, 1991). Biases do no longer demonstrate themselves when these are taken into account (Gigerenzer, 1991).

Even though the heuristics and biases approach may not in all cases describe human decision-making correctly, Tversky and Kahneman's early work has influenced other theories of human decision-making and judgment, one of them being a notable family of theories called *dual-process theories*. Tversky and Kahneman (1974; 1983) and Epstein (1994) had a similar view that there are two different modes of reasoning (Osman, 2004). Epstein (1994) labelled them as "a rational system" and "an emotionally driven experiential system", whereas Tversky and Kahneman (1983) called them *intuitive* and *extensional*. The rational, or extensional information processing, is conscious, slow, logical, and detail-oriented (Epstein, 1994). Experiential or intuitive (or natural, as Tversky and Kahneman also describe it) is automatic, affective, faster, and deals with more broader conceptions (Epstein, 1994). Intuitive judgment of probability relies on heuristics (Tversky & Kahneman, 1974), and is carried out by the experiential thinking system.

There are three theories of dual-reasoning that have had the most significant impact on developing the research on reasoning behaviour, judgment, decision-making and problem solving: Evans and Over's dual-process theory, Sloman's two-system theory and Stanovich and West's two-systems theory (Osman, 2004). In 1996, Evans and Over introduced their dual process theory, which integrated earlier proposals of dual-processing into one theory (Evans, 2002). The theory suggests that two separate cognitive systems make the foundations of implicit and explicit cognitions (Evans & Over, 1996). These systems were later named System 1 and System 2 by Stanovich (1999) (Evans & Stanovich, 2013). Stanovich's research on individual differences of reasoning provided empirical evidence for the existence of two different reasoning systems (Evans & Stanovich, 2013). About the same time, Sloman (1996) proposed a relatively similar theory that distinguished associative from rule-based reasoning.

Behind the Evans and Over's dual-process theory is the Evans' heuristic-analytic -theory (Osman, 2004). According to the theory, there are two different reasoning processes: pre-attentive *heuristic process* that is followed by an *analytical process* (Evans, 1984). The heuristic process selects relevant information for the analytic process (Evans, 1984). Information judged as irrelevant is not processed any further (Evans, 1984). Evans (1984) claims that without understanding heuristic processes, human rationality remains unclear. It is important to notice that Evans defines the term 'heuristic' very differently from Tversky and Kahneman (1974). In Evans' work (1984), a heuristic refers to pre-attentive processes which select relevant information for analytic process. In order to find out whether reasoning is rational, it should be known *what* is reasoned about, and that is what the heuristic process determines (Evans, 1984). The purpose of analytic processes is to evaluate the selected information and make a judgment, inference or a decision based on it (Evans, 1984). Analytic processes represent deliberate, explicit thinking, that is context-dependent in a sense that the individual's experience has an effect on it (Evans, 1984; Osman, 2004). They still "accomplish to some forms of logical analysis", though (Osman, 2004).

Evans and Over's work from 1996 defines the System 1 as an implicit system, that is pragmatic, associative, rapid, high-capacity, efficient and driven by past learning. System 2 is slow and sequential, capable of logical problem solving and hypothetical thinking (Evans, 1996). The capacity of working memory limits the functioning of System 2 (Evans, 1996). Though first labelled as System 1 and System 2, both Evans and Stanovich have discontinued using the terms, and instead describe the two different types of reasoning Type 1 and Type 2 (Evans & Stanovich, 2013). Evans and Stanovich (2013) claim that based on the literature, the key feature of Type 1 processing is autonomous. The ability to preserve the memory trace of a stimulus in mind and manipulate or work with its mental representation defines the Type 2 processing (Evans & Stanovich, 2013). Type 2 thinking correlates with the measure of general intelligence, and it can override Type 1 thinking if necessary (Evans & Stanovich, 2013).

Fuzzy-trace theory (FTT) is a dual-process theory of memory and reasoning, that can be applied to explain and predict behaviour and decision-making (Brainerd & Reyna, 2001; Helm et al., 2017). The theory posits that there are two types of memory representation: verbatim and gist (Reyna et al., 2015; Helm et al., 2017). Verbatim representations are exactly memorised information (Brainerd & Reyna, 1990). Gist representations are intuitive and fuzzy – they represent the underlying meaning of the information (Helm et al., 2017; Brainerd & Reyna, 1990). According to FTT, verbatim and gist processing both develop with age, but people are more likely to rely on gist processing as their development proceeds, and also when their expertise in some area increases (Reyna et al., 2014; 2015). Expert decision-making has some unique features. Experts are proven to make better decisions within their domain of expertise comparing to novices (Helm et al., 2017).

Still, experts are prone to be affected by the same cognitive biases than everyone else, as Tversky and Kahneman proved in 1974 (Helm et al., 2017). Even statistically sophisticated individuals are exposed to the same fallacies than people who have no education in statistics when dealing with more complex problems (Tversky & Kahneman, 1974). Common heuristics (representativeness, availability, and anchoring) lead to systematic errors in judgments and decision-making under uncertainty (Tversky & Kahneman, 1974). Instead of following statistical rules of reasoning, people violate the assumptions of rationality (Kahneman & Tversky, 1974). A study made by Reyna et al. (2014) demonstrated that experienced intelligence agents "exhibited larger decision biases than college students". According to the experiment, expert intelligence agents were more ready to risk human lives when outcomes were framed as losses rather than gains (Reyna et al., 2014).

In traditional dual-process theories, intuition is in contrast with reasoning (Reyna, 2013). But in FTT, intuition is not only part of reasoning, it is "the default mode of adult reasoning, that generally determines judgments and decisions" (Reyna, 2013). In FTT, intuition and impulsivity are not related processes (Reyna et al., 2015; Reyna, 2013). Experts can make fast, insightfully intuitive inferences because they process and retrieve gist information rapidly (Reyna, 2013). Traditional dual-process theories and heuristics and biases approach think that when people rely on intuitive, fast, and automatic thinking processes and analyse the information by using heuristics, they often make errors and do not think as rationally as they could (Helm et al., 2017; Reyna, 2013; Reyna et al., 2015). This has been the conclusion, even though Tversky and Kahneman (1974) themselves noted that heuristics are useful and there is a reason they are used. FTT posits a contrary point of view. It is acknowledged that expert decisions are biased as well and relying on gist information can also lead to biases (Helm et al., 2017; Reyna, 2013). But at the same time, intuition and fast processing is not inherently deemed as an unreasonable way to think. Actually Helm et al. (2017) state that "gist-based processing is - - developmentally advanced", because it emerges later in life and grows along the expertise, and because it allows reflecting meaningful distinctions that actually matter when making decisions.

In the context of multiobjective optimisation, FTT is a considerable theory in explaining how humans and especially experts make decisions. The DM in multiobjective optimisation is being regarded as an expert regarding the multiobjective optimisation problem to be solved, as defined in Chapter 2.1. If the assumptions of FTT are true, it could be beneficial to support the DM's gist processing. This could result in biases, but it could also result in better decisions. In FTT, reliance on gist processing is associated with framing bias and hindsight bias (Helm et al., 2017, cf. e.g., Reyna, 2013; Reyna, 2005). At least it should be acknowledged that experts might process information differently and rely on fuzzy representations of information. Regardless of its intuitive nature, gist processing is not cognitively less demanding. The impact of cognitive load in the context of interactive multiobjective optimisation is addressed in the following section.

3.2 Decision-makers interacting with decision-making systems

Even though interactive multiobjective optimisations methods of the DESDEO software framework are operated on a user interface, this approach is not common. The research contribution of this thesis is to develop a scale or scales to assess and compare interactive multiobjective optimisation methods. The results of this study can further the design and research of interactive multiobjective optimisation methods when researchers and practitioners can rely on reliable, validated measurement scales to test these methods.

When decision-makers use interactive multiobjective optimisation methods of DESDEO to solve optimisation problems, they also interact with the decision-making system. In addition, for understanding human decision-making and judgment, it is beneficial to examine factors that affect HCI. HCI is an interdisciplinary field of computer science and psychology which pursues to understand and support humans interacting with and through technology (Carroll, 1997). One of its missions is to understand the detailed involvement of cognitive, perceptual, and motor components in the interaction between a human and a computer (Olson & Olson, 2003). Carroll (1997) defines HCI as a science of design. One of the goals of HCI is to design useful and usable technology (Olson & Olson, 2003).

3.2.1 Usability and satisfaction

Usability is a fundamental concept in HCI (Hartson, 1998; Hornbæk, 2006). Defining usability is not simple and there has been disagreement in HCI about what usability means (Lewis, 2014; Tractinsky, 2018). Part of the difficulties of defining the term, according to Lewis (2014), stems from the fact that measuring usability is a complex task. Usability is not a specific property of a product, service or a system (Lewis, 2014). This had been acknowledged in the previous ISO standard from 1998, that emphasized usability as a result of interaction (Bevan et al., 2015).

The revised version of ISO 9241-11:2018 defines usability as "extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" (International Organization for Standardization, 2018). The previous standard from 1998 has been withdrawn and it was, for the most part, almost identical with the current one (Bevan et al., 2015). According to the Bevan et al. (2015), ISO 9241-11:1998 became internationally recognised as a foundation for understanding and applying usability. In usability engineering, the definition of usability as a measurement of effectiveness, efficiency, and satisfaction, is relatively well adopted in current practices (Lewis, 2014).

How to improve the usability of interactive systems is an important research question in HCI (Hornbæk, 2006). Research on this topic has generated ways to test, improve, and measure usability (Hornbæk, 2006). A common method to spot usability problems in a product, system or a service is to perform a heuristic usability evaluation (Quiñones & Rusu, 2017). Nielsen's 10 heuristics are well recognized and widely accepted in the field (Quiñones & Rusu, 2017; Jimenez et al., 2016). Nielsen calls them "heuristics" because they should be interpreted as general guidelines for usability, not as specific rules (Nielsen, 2005). These heuristics do not encompass all the elements and features of usability in every domain (Quiñones & Rusu, 2017; Jimenez, Lozada & Rosas, 2016). Usability can also mean different things to novices versus experts: for novice and casual users, usability is more often defined as understandability or learnability, whereas experts expect usefulness and functionality (Hartson, 1998). In both cases, usability is defined by measures of productivity, performance and satisfaction (Hartson, 1998). Users just have different criteria for these metrics (Hartson, 1998). Evaluating and measuring usability is therefore highly context-dependent (Lewis, 2014).

The concept of *satisfaction* is a part of the usability standard ISO 9241-11:2018 (International Organization for Standardization, 2018) and an important aspect of usability in HCI research (Hartson, 1998; Hornbæk, 2006; Lewis, 2014), although sometimes it is simply being viewed as a side product of effectiveness and efficiency (Bevan, 2010). The ISO 9241-11 version from 1998 defined the satisfaction as "freedom from discomfort, and positive attitudes towards the use of the product" (Bevan et al., 2015; Bevan, 2010). The current standard defines satisfaction as "extent to which the user's physical, cognitive and emotional responses that result from the use of a system, product or service meet the user's needs and expectations" (International Organization for Standardization, 2018).

Hornbæk (2006) reviewed 180 studies and analysed how usability was measured. Satisfaction was measured in 112 studies, but only few papers mentioned the actual questions they used in those questionnaires (Hornbæk, 2006). Standardized questionnaires were rarely used (Hornbæk, 2006). When standardized questionnaires were not used, users were asked about ease-of-use, their attitudes towards the interface and its content, their perception of outcomes and the process of interaction, and few other measures like beauty (Hornbæk, 2006). *Perception of outcomes* is about how users view the outcomes of the interaction (Hornbæk, 2006). This is measured by asking users to assess their own performance, understanding and learning, and to evaluate the success of the task outcome and their confidence about it (Hornbæk, 2006). *Perception of interaction* is about how users see the process of interaction (Hornbæk, 2006). Often it means users' perception of task complexity and task completion times (Hornbæk, 2006). In addition to these, satisfaction was measured by a large amount of other indicators, such as: users' sense of control, experience of engagement, experience of physical discomfort, attitude towards using the interface again, connection to other persons or to interface, feelings of presence, satisfaction with specific features in the specific context of use, quality of the information, and so on (Hornbæk, 2006). As Hornbæk (2006) explains the situation, "the measurement of satisfaction seems in a state of disarray". Even though there are standardised questionnaires, a lot of the studies measure satisfaction by their own ways.

The Questionnaire for User Interface Satisfaction (QUIS) is a measurement tool, that measures the user's subjective rating of the human-computer interface

(Chin et al., 1988). According to Chin et al. (1988), user acceptance of a system, or in other words, subjective satisfaction, is a critical measure of a system's success. A system's overall performance can be evaluated to be good, but if the user is dissatisfied with the system and its interface, they may not be willing to use it again (Chin et al., 1988). QUIS includes 27 items in five categories: overall reactions to the systems, characters on the screen, terminology and system information, learning to use the system, and system capabilities like correcting mistakes and taking into account different users (Chin et al., 1988). The user is asked to evaluate how they perceive the properties of the system, like rating the position of messages on screen inconsistent – consistent, and their personal experience, like overall reactions rating between terrible – wonderful, frustrating – satisfying, and so on (Chin et al., 1988).

ASQ (After-Scenario Questionnaire) is a validated measurement scale to measure user satisfaction with system usability (Lewis, 1995). The items address the ease of task completion, time to complete the task, and adequacy of the support information (Lewis, 1995). The user is asked to evaluate how satisfied they are in each of these aspects (Lewis, 1995). Questionnaires to measure satisfaction published between 1974 and 1988 were reviewed in LaLomia and Sidowski (1990). The first standardized usability questionnaires suitable for usability testing were QUIS by Chin et al. (1988), ASQ by Lewis (1995), Post-Study System Usability Questionnaire (PSSUQ) by Lewis (1992), and Computer User Satisfaction Inventory (CUSI) by Kirakowski and Dillon (1988) (Lewis, 2014). Nowadays most often used standardized usability questionnaires are QUIS, the Software Usability Measurement Inventory, PSSUQ, and the System Usability Scale (SUS) (Lewis, 2014).

User experience (UX) is a central concept in the field of HCI, even though there is not a clear consensus of its definition and extent (Law et al., 2009; Lallemand et al., 2015; Hassenzahl, 2003). UX is sometimes seen as a part of more traditional concept of usability, but it is also recognized to have unique features (Lallemand et al., 2015). Usability is more focused on task performance, whereas UX research is more interested in unique, lived experiences, and especially hedonic and enjoyable elements of interaction (Hassenzahl, 2003; Vermeeren et al., 2010; Bargas-Avila & Hornbæk, 2011; Lallemand et al., 2015). Vermeeren et al. (2010) note that satisfaction, the subjective component of usability, can also be seen as a part of UX evaluation, but it is nowhere near the only subjective quality that the UX research is interested in. Satisfaction has been linked to the hedonic consequences of UX models (Hassenzahl, 2003), and in some studies usability (e.g., effectiveness, efficiency, satisfaction) has been measured along with the UX (Bargas-Avila & Hornbæk, 2011). Usability and UX are not completely distinguishable, because usability affects the quality of the user experience, and components of user experience can also influence how the usability of a system is perceived (Sharp et al., 2007, cited in Hollender et al., 2010).

Hollender et al. (2010) refer to the study made by Kurosu and Kashimura (1995), which results were later confirmed by Tractinsky (1997), that perceived aesthetics attractiveness of a system has a strong influence on its apparent usability.

To conclude, usability and satisfaction are strongly connected to each other. Satisfaction is often measured by assessing the rate of learnability, ease of use, general feelings towards the system, willingness to use the system again, how helpful, relevant or consistent the elements of the system are perceived, and so many more. This demonstrates the vagueness and multidimensionality of the concept of satisfaction. Satisfaction is also studied in UX, but it is not the most central concept in it. In usability measurement, satisfaction is associated with task and user performance, which are very different things than the subjective notions of satisfaction in UX research.

3.2.2 Cognitive load

Cognitive load theory (CLT) has a lot of influence and is widely acknowledged within the fields of instruction, learning and educational psychology (Schnotz & Kürschner, 2007; Martin, 2014). Today, it influences especially the fields of digital technologies and e-learning (Martin, 2014). CLT, initially developed in the 1980's, is based on schema theory and the idea of limited human short-term memory (Hollender et al., 2010). Its original purpose was to describe instructional design of pedagogical methods (Kosch et al., 2023).

Short-term memory, or *working memory*, is a system that temporarily maintains and manipulates information necessary to perform complex cognitive tasks, such as learning and reasoning (Baddeley, 1992). Its capacity is limited and can only process a certain amount of information at the given moment (Miller, 1956). During the time in which the information resides in the working memory, it gets transferred to the long-term memory (Atkinson & Shiffrin, 1968). *Long-term memory* is a relatively permanent storage (Atkinson & Shiffrin, 1968), capable of holding almost unlimited amounts of information (Sweller, 1994; Sweller et al., 1998). *Schema* is a cognitive construct that encompasses multiple elements of information, organized as one (Sweller, 1994; Sweller, 2005). Sweller (1994) illustrates how schemas work by presenting an example of a tree schema: instead of dealing with each element of a tree (leaves, branches, colour) individually in memory, a person relies on a tree schema that contains all the elements associated with a tree in their mind. Automated schemas are held in the long-term memory (Sweller, 2005). Schemas allow humans to process complex material that exceeds the capacity of their short-term memory, because they only need to process one element of information (Paas, Renkl & Sweller, 2003). Learning happens when schemas are constructed (Sweller et al., 1998; Sweller, 2005). According to Martin (2014), CLT claims that the ease or difficulty of learning a specific material depends on how well people are able to process the information that is needed to solve a problem or learn something new by retrieving or constructing a schema.

The primary concern of CLT is how easily information is processed in working memory (Sweller et al., 1998). In CLT theory, there are three types of cognitive load that affect working memory load: intrinsic cognitive load, extraneous cognitive load, and germane cognitive load (Sweller et al., 1998). *Intrinsic cognitive* *load* is the inherent complexity of the task or learning material itself (Sweller et al., 1998). It is affected by the qualities of the task or material at hand and the learner's level of expertise (Sweller et al., 1998). Therefore, it is not possible to modify the amount of intrinsic cognitive load by making changes in instructional design (Sweller et al., 1998). *Extraneous cognitive load* is the result of poorly designed instructions (Sweller et al., 1998). They way in which information is presented affects extraneous cognitive load, and therefore it is possible to reduce it by instructional intervention (Sweller et al., 1998; Sweller, 1994). *Germane cognitive load* signifies the effort invested in constructing schemas (Sweller et al., 1998). (Sweller et al., 1998). At appropriate levels, germane cognitive load supports learning (Sweller et al., 1998). Good instructional design decreases extraneous cognitive load but increases germane cognitive load (Sweller et al., 1998). Total cognitive load is the sum of three cognitive load types (Sweller et al., 1998). Learners can be encouraged to engage in material that increases germane cognitive load to help benefit their learning, but it only works if the total cognitive load stays within the working memory limits (Sweller et al., 1998).

Hollender et al. (2010) carried out a literature review with a goal to integrate the concepts of CLT and HCI in order to advance the design and research of e-learning environments. They discovered that both fields share the same underlying beliefs regarding the human cognitive system, focus on reducing irrelevant cognitive load, and have recently adopted a notion that sometimes cognitive load can be beneficial for the learning (Hollender et al., 2010). The HCI research adopted the notion of limited human working memory and today it is common knowledge in research and practice, including design heuristics (Rogers, 2004). Reducing the user's cognitive load as much as possible became an important goal in interaction design (Mandel, 1997; Preece et al., 2002, cited in Hollender et al., 2010). In software design, instructional design principles of CLT have been applied in a similar way than usability goals and principles (Hollender et al., 2010).

In HCI, terms of "cognitive workload", "cognitive load", and "mental workload" are all used to describe the cognitive demand of a task (Kosch et al., 2023). The concept of mental workload refers directly to the cognitive demands of the task itself (Miyake, 2001). The concept of CLT (i.e., intrinsic, extraneous and germane cognitive load) is important to understand as its own distinct theory. In practice, however, these concepts often get mixed up and are sometimes used interchangeably (Kosch et al., 2023; Hollender et al., 2010). Kosch et al. (2023) use the term "cognitive workload" to describe "workload imposed through the instructional system design of user interface visualizations or cognitive demand of users who process information" (Kosch et al., 2023). Regardless of the term confusion, it is apparent that the influence of CLT is evident in measuring and understanding cognitive load in HCI research and practice (Hollender et al., 2010). Usually in HCI, cognitive load is measured by asking participants to fill a questionnaire after completing the task (Kosch et al., 2023). NASA Task Load Index (NASA-TLX), originally developed to study the task load of pilots, is the most often used questionnaire to measure participants' perceived subjective cognitive load (Hart & Staveland, 1988; Hart, 2006; Kosch et al., 2023). NASA-TLX is easy to use and that may be the reason for its popularity (Kosch et al., 2023).

Learning is strongly embedded in the concept of CLT. Learning processes have a role in HCI as well in two ways: if a novice user wants to complete a specific task on a computer system, they first need to learn to use it (Nielsen, 1994, cited in Hollender et al., 2010). The other aspect of learning in HCI are educational software that support the user in obtaining knowledge and skills in different areas of interest (Hollender et al., 2010). Interactive multiobjective optimisation methods of the DESDEO software framework, or other interactive multiobjective optimisation methods, are not educational tools per se. However, learning is an inherent part of interactive multiobjective optimisation, and the methods designed for that also support the DM's learning process. When the DM uses an interactive multiobjective optimisation method of the DESDEO software framework, they learn how to use the software the method has been implemented on, to use the method itself, and about the problem to be solved and concepts of multiobjective optimisation altogether. Therefore, the concept of cognitive load and more specifically, the concept of CLT, could help our understanding of the learning process in interactive multiobjective optimisation. In a lot of the HCI research, low cognitive load is often interpreted as an indicator of high system quality without explicating what those features are that constitute to a users' perception of a quality (Kosch et al., 2023). Usually, the goal is to decrease cognitive load as much as possible, even though germane cognitive load can support schema construction and therefore learning. Fostering germane cognitive load contradicts with the basic principles of usability and it is therefore not so utilised in HCI research and practice (Hollender et al, 2010).

The study this thesis is based on, implements the modified NASA-TLX to measure cognitive load, along with two other questionnaire items that measure the feeling of tiredness and whether the DM feels that the number of iterations to reach the acceptable solution was too high (Afsar et al., 2024). Most beneficial in the context of interactive multiobjective optimisation, especially regarding the importance of learning, would be to measure cognitive load with a measurement instrument that includes the different cognitive load types (intrinsic, extraneous and germane), as explained above. However, using NASA-TLX to measure cognitive load is helpful to understand what kind of impact the user interface of the given interactive multiobjective optimisation method has on users' perceived cognitive load.

3.3 Desirable properties of interactive multiobjective optimisation methods

Afsar et al. (2021) conducted a literature review where they collected assessments and comparisons of interactive multiobjective optimisation methods from 45 papers. Based on their expertise and the literature review, Afsar et al. (2021) have proposed a set of desirable properties of interactive multiobjective optimisation methods. Learning and decision phase have been separated from each other and they both have their own desirable properties. There are also general properties that are present in the solution process regardless of its phase. The purpose of the study by Afsar et al. (2021) is to investigate how the performance of interactive methods has been assessed in the published literature.

When solving multiobjective optimisation problems, the solution process can often be observed to contain two phases, a learning phase and a decision phase, that have different objectives (Miettinen et al., 2008). In the *learning phase*, the DM learns about the qualities and characteristics of the problem so that they could understand what kind of solutions are possible (Eskelinen et al., 2010; Miettinen et al., 2008). Once the DM has identified a region of interest, they can proceed to the *decision phase*, where they refine the search in the region of interest to find the MPS (Miettinen et al., 2008). The learning phase and the decision phase may be used iteratively in interactive multiobjective optimisation methods (Miettinen et al., 2008).

As a result of study by Afsar et al. (2021) following desirable properties, presented in Table 1, were proposed:

TABLE 1 Desirable properties

General properties (GP)	GP1 - The method captures the preferences of the DM. GP2 - The method sets as low cognitive burden on the DM as possible. GP3 - A user interface supports the DM in problem solving. GP4 - The DM feels being in control while interacting with the method. GP5 - The method prevents premature termination of the overall solution process.
Learning phase (LP)	LP1 - The method helps the DM avoid anchoring. LP2 - The method allows exploring any part of the Pareto optimal set. LP3 - The method easily changes the area explored as a response to a change in the preference information given by the DM. LP4 - 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. LP5 - The method properly handles uncertainty of the information provided by the DM. LP6 - The method allows the DM to find one's region of interest at the end of the learning phase.
Decision phase (DP)	DP1 - 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.

	DP2 - The method reaches the DM's MPS. DP3 - The method allows the DM to fine-tune solutions in a reasonable number of iterations and/or reasonable waiting time. DP4 - The method does not miss any Pareto optimal solution that is more preferred (with a given tolerance) for the DM than the one chosen.

It is important to assess the performance and qualities of interactive methods because some methods are more suitable for one of the phases (Afsar et al., 2021). For example, tradeoff-free methods like E-NAUTILUS are more appropriate for the learning phase (e.g., Miettinen et al., 2010; Afsar et al., 2024), and classification-based methods like NIMBUS are more fitted for the purposes of the decision phase, like reported in Afsar et al. (2024). If interactive methods could be evaluated, it would help to determine which method to use in solving a particular optimisation problem (Afsar et al., 2021).

These properties have been studied in two empirical research: Afsar et al. (2023) and Afsar et al. (2024). Six desirable properties were selected to be studied in Afsar et al. (2023). The desirable properties were connected to specific research questions, and again operationalised into questionnaire items (Afsar et al., 2023). However, the results were not interpreted against the desirable properties (Afsar et al., 2023). The study made by Afsar et al. (2024) followed a similar procedure. They selected eight desirable properties, which were connected to three research questions and then operationalised into questionnaire items (Afsar et al., 2024).

4 METHOD

This chapter describes the methodological position of this thesis, the study it is based on, and the research methods used. Because the research data is gathered in a study made by Afsar et al. (2024), especial attention is paid to the reasoning behind the operationalisations of the given study in Section 4.3.

4.1 Methodological position

Wobbrock and Kientz (2016) introduce seven research contribution types in the field of HCI. The contributions of this thesis are mostly methodological, since its purpose is to develop a new reliable way to measure previously identified phenomena (Wobbrock & Kientz, 2016). The results of this thesis intent to improve future research and design of interactive multiobjective optimisation methods.

Jokinen (2015) has proposed the methodological framework for the research of human-technology interaction (HTI). There are four methodological positions: behaviourism (*empiricism*), neuroscience (*physicalism*), subjectivism (*phenomenology*), and cognitivism (*functionalism*) (Jokinen, 2015). These are presented in Figure 1.

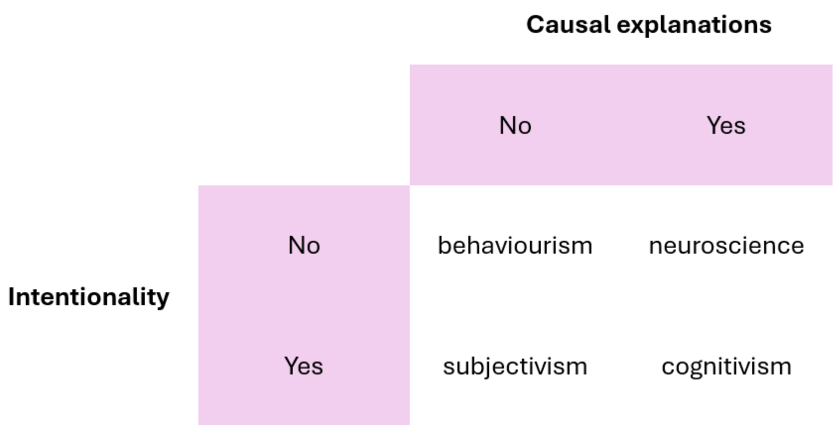

FIGURE 1 Methodological framework of HTI research by Jokinen (2015)

It is not practical to try to fit every study into one of these positions, since studies can combine aspects from different methodological positions and do not necessarily fit perfectly to any of them (Jokinen, 2015). But the framework is helpful when evaluating the assumptions behind the studies of HTI (Jokinen, 2015).

Behaviourism rejects the notions of causality and intentionality (Jokinen, 2015, Figure 1). In behaviourism it is thought that inferences about human thinking cannot be made, and psychology should focus on studying objectively measurable events, such as behaviour (Watson, 1913). Neuroscience is the study of the brain and the nervous system. It views the human mind from a physical perspective, allowing causality but denying the assumption of intentionality (Jokinen, 2015, Figure 1). Subjectivism studies the human experience (Jokinen, 2015). Intentionality is assumed but causal explanations are not, because it would be hard to generate only one particular mental cause and distinguish it from others invoked (Jokinen, 2015). In cognitivism, cognitive processes direct human behaviour (Haugeland, 1978). Methodologically it accepts the assumption of intentionality and causality (Jokinen, 2015, Figure 1). It is assumed that it is possible to examine the effect that mental states have on human behaviour (Jokinen, 2015).

The study reported in this thesis follows loosely the subjectivist and cognitivist school of thoughts. It is assumed that people have inner mental states, experiences, and mental representations that affect their behaviour and that these can be examined by conducting an experiment.

4.2 Validity and reliability

The goal of developing a scale is to create a valid measure of an underlying construct (Clark & Watson, 1995). Previous stages of scale development, like conceptualization and literature review, have been conducted by Afsar et al. (2021, 2024). This research focuses on examining the decisions made on previous studies and aims at establishing a reliable scale based on their data via statistical data analysis.

Validity describes the accuracy of a measurement scale or a survey (Litwin & Fink, 1995). More specifically, it indicates how well the measurement instrument actually assesses the construct that it is supposed to assess (Nunnally & Bernstein, 1994; Litwin & Fink, 1995). Construct validity is an important concept in psychological research (Westen & Rosenthal, 2003). According to Westen and Rosenthal (2003), construct validity describes how well the variance in the measured scores or results represent the variations in the underlying construct that the measurement instrument is supposed to measure. Construct validity is better understood as both continuous and continual (Westen & Rosenthal, 2003). The former means that validity is a non-binary concept, a matter of degree (Westen & Rosenthal, 2003). Continual refers to the idea that validity assessment is an ongoing, self-refining process, which often requires re-evaluation of measure and construct (Westen & Rosenthal, 2003).

Reliability is a statistical measure which indicates the degree of reproducibility of the survey instrument's data (Litwin & Fink, 1995). There are three common methods to assess the reliability of a survey instrument: test-retest, alternate-form, and internal consistency (Litwin & Fink, 1995). Test reliability can also be assessed by commonly used Cronbach's alpha (Tavakol & Dennick, 2011). Cronbach's alpha, developed by Lee Cronbach (1951), is a measure of the internal consistency of a test or a scale. Internal consistency represents how well the test items measure the same construct or concept (Tavakol & Dennick, 2011). Before the scale can be applied in future research, internal consistency should be determined to ensure its reliability (Tavakol & Dennick, 2011).

The acceptable value of alpha is between .70 and .90 (Nunnally & Bernstein, 1994). The value of Cronbach's alpha increases when the items in a test are intercorrelated (Tavakol & Dennick, 2011). Simply selecting highly intercorrelated items to increase the value of alpha and therefore the measured internal consistency of the test, is not, however, advisable. After a certain level, increased internal consistency does not improve the construct validity, but can in fact lower it (Clark & Watson, 1995). That is because highly intercorrelated items are redundant: together they do not offer any more information about the underlying construct than just selecting any one of them (Clark & Watson, 1995). Therefore, to ensure the internal consistency of the scale, its items should only moderately intercorrelate (Clark & Watson, 1995).

4.3 Operationalisation

Operationalisation enables measuring a concept or a term for the purposes of a specific research (Niiniluoto, 1980, s. 187). It gives the concept or the term a measurable indicator (Niiniluoto, 1980, s. 187), like questionnaire items measured with a Likert-scale which assigns them a numerical value. The indicator should measure the concept or the term as well as possible, without measuring something else (Niiniluoto, 1980, s. 187). Operationalisation impacts the validity of a research and a measurement instrument (Niiniluoto, 1980, s. 187).

In the study by Afsar et al. (2024) a survey was conducted where questionnaire items for comparing interactive methods were based on the previously identified desirable properties of interactive methods, proposed in Afsar et al. (2021). There were three research questions: how extensive the cognitive load of the whole solution process (cognitive load) is, how well the method captures and responds to the DM's preferences (capturing preferences and responsiveness), and is the DM satisfied with the overall solution process and confident with the final solution (satisfaction and confidence) (Afsar et al., 2024). Nine desirable properties were selected altogether and connected to research questions (Afsar et al., 2024).

In Table 1, the research questions, corresponding desirable properties and questionnaire items that measure the properties are listed. The next three subsections cover the properties and how they were operationalised and if the reasoning behind these operationalisations is sound. Only Likert-scale items were addressed in these discussions. There is a plus-sign at the end of the item if it was both Likert-scale and open-ended.

TABLE 2 Operationalisations

RQ	**Desirable properties**	**Questionnaire items**
Cognitive load	The method sets as low cognitive burden on the DM as possible.	I am satisfied with my performance in finding the final solution (+) A lot of mental activity was required (e.g., thinking, deciding, and remembering). I had to work hard to find the final solution. I felt frustrated in the solution process (e.g., insecure, discouraged, irritated, stressed).
	The method allows the DM to fine-tune solutions in a reasonable number of iterations and/or reasonable waiting time.	It took too many iterations to arrive to the acceptable solution. I felt tired.
Capturing preferences and responsiveness	The method captures the preferences of the DM.	The preference information was easy to provide. I was able to reflect my actual preferences when providing the information required by the method (+)
	The DM feels being in control while interacting with the method.	It was easy to learn to use this method. I felt I was in control during the solution process. I felt comfortable using this interactive method. The method has all the necessary functionalities. I was able to return to previous solutions whenever I needed in the solution process.
	The method easily changes the area explored as a response to a change in the preference information given by the DM.	The solution(s) I obtained reflects my preference information well.

		It was easy to explore solutions with different conflicting values of the objective functions. In general, the method reacted to the preference information I provided.
Satisfaction and confidence	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.	After this iteration, I know more about the problem. I think that the solution I found is the best one. I obtained a clear idea of the values that the objectives can simultaneously achieve. I obtained a clear idea of the possible choices available similar to the solutions I was interested in.
	The method does not miss any Pareto optimal solution that is more preferred (with a given tolerance) for the DM than the one chosen.	I am satisfied with the solution I chose(+)
	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.	I am satisfied with the final solution. Overall, I am satisfied with the ease of completing this task. Overall, I am satisfied with the amount of time it took to complete this task. Overall, I am satisfied with the support information (online help, messages, documentation) when completing this task.

4.3.1 Cognitive load

Cognitive load was assessed by six items. Four of them were assigned to the property “The method sets as low cognitive burden on the DM as possible”. These items are based on NASA-TLX-questionnaire (Hart & Staveland, 1988) with little modifications. NASA-TLX has six subsections which are mental, physical, and temporal demands, performance, effort, and frustration (Hart, 2006). Items in the given study took into account mental demands, performance, effort, and frustration. Temporal and physical demands in context of this study were not relevant (Afsar et al., 2024).

The property “The method allows the DM to fine-tune solutions in a reasonable number of iterations and/or reasonable waiting time” was also seen as part of a cognitive load assessment (Afsar et al., 2024). The property had been operationalised to the following items:

- It took too many iterations to arrive to the acceptable solution.
- I felt tired.

This property has been categorized into the decision phase in Afsar et al. (2021). After the learning phase, the DM has identified the region of interest that best seems to fit their preferences (Afsar et al., 2021). In the decision phase, the DM refines the search within this region of interest until the MPS is found (Afsar et al., 2021). Afsar et al. (2021) claim that since the DM's chosen region of interest is based on what they learned about the problem in the learning phase, the DM should be sure about the region of interest at the beginning of the decision phase. Because of that, refining the solution until the DM finds the MPS, should not take the method too long or too many iterations (Afsar et al., 2021).

It can be argued that some elements of multiobjective optimisation are cognitively difficult. Long-lasting, cognitively demanding tasks can lead to mental fatigue, impacting task performance (Van der Linden et al., 2003; Borragán et al., 2017). Maintaining high cognitive performance can lead to mental fatigue (Hart, 2006). Therefore, investigating the DM's feeling about tiredness and their opinion about the number of iterations is a reasonable way to measure cognitive load. NASA-TLX ratings have also been studied together with other factors, including fatigue (Hart, 2006).

4.3.2 Capturing preferences and responsiveness

During the learning process, the DM explores the given solutions based on their preferences, which causes variation in solutions between iterations (Afsar et al., 2024). Because of this, it is important to assess the method's ability to react and respond to preference information given by the DM (Afsar et al., 2024). The ability can be measured by assessing how easy it is for the user to give the preference information, whether the method has all the necessary functionalities for the DM to do that, and if the DM feels that they control the solution process (Afsar et al., 2024).

This research question had three properties assigned to it. The property "The method captures the preferences of the DM" was measured by two items:

- The preference information was easy to provide.
- I was able to reflect my actual preferences when providing the information required by the method.

The purpose of the first item is to evaluate how well the DM is able to express their preferences during the solution process (Afsar et al., 2024). The second item assesses how easy the method makes it for the DM to provide their preferences (Afsar et al., 2024). For the sake of obtaining the favourable solution and being able to use the method effortlessly, it is crucial that the DM succeeds in providing their preferences. That is because interactive multiobjective optimisation methods only generate Pareto optimal solutions that interest the decision maker (Thiele et al., 2009). The method's ability to capture the preferences and respond to them can therefore be measured by these two items, because they assess how well the method fulfils its purpose, that is to help the DM to understand the problem and their own preferences.

The property "The DM feels being in control while interacting with the method" was operationalised to five items:

- It was easy to learn to use this method.
- I felt I was in control during the solution process.
- I felt comfortable using this interactive method.
- The method has all the necessary functionalities.
- I was able to return to previous solutions whenever I needed in the solution process.

The DM's feeling of control is a fitting way to assess the method's ability to capture preferences and respond to them. That is because the solutions generated by the methods are (or they should be) based on the DM's preference information. That means that the DM should control the solution process: proposed solutions are not random, but directly reflect the DM's preferences. If the DM feels that using the method is hard, annoying, and unpleasant, and difficult to learn, the DM does not feel that they control the process, and therefore, they cannot be sure that the method reflects their actual preferences.

The property "The method easily changes the area explored as a response to a change in the preference information given by the DM" was assessed by three questionnaire items:

- The solution(s) I obtained reflects my preference information well.
- It was easy to explore solutions with different conflicting values of the objective functions.
- In general, the method reacted to the preference information I provided.

It is important to note that this property is assigned to the learning phase in Afsar et al. (2021). In the learning phase, the DM gets to know and learn about the set of Pareto optimal solutions, tradeoffs among the objective functions, areas with different conflict degrees and values of the decision variables in each area (Afsar et al., 2021). The first item assessing this property is intended to measure the method's ability to reflect the DM's preference information even when the explored area changes. Some methods adapt to changing preference information better than the others, and if the method cannot do it well, it probably does not reflect the DM's preferences well anymore. This item could, however, address the property "The method captures the preferences of the DM" better. Afsar et al. (2021) also assigned this property to general properties, not specifically to the learning or decision phase. The second item assesses the method's responsiveness as the ability to react to the DM's preferences (Afsar et al., 2024). As mentioned above, the method should react to changes in preference information, even to major ones. Regardless of the solutions changing, the DM should still be able to easily get a good overall comprehension of offered solutions and how they differ.

4.3.3 Satisfaction and confidence

The third research question is about the DM's satisfaction with the overall process and the final choice they made, and whether they are confident that the choice they made was right (Afsar et al., 2024). Afsar et al. (2024) argue that the DM's satisfaction with the overall solution process and their satisfaction and confidence with the final solution should be distinguished from each other. That is because even though the final solution pleases the DM, they might still find the interactive solution process cognitively too demanding or just hard to understand, therefore affecting their willingness to interact with the method again (Afsar et al., 2024).

Being involved in the decision process increases the DM's confidence towards the final solution in general (Miettinen, 1999). A reason could be that while interacting with the method during the solution process, the DM simultaneously learns more about the problem. Belton et al. (2008) argue that a willingness to participate in the process again means that the DM is satisfied with the method. The satisfaction is possibly related to learning as well (Belton et al., 2008).

The third research question is connected with three desirable properties that were operationalised to nine (Likert-scale) items. The properties linked to satisfaction and confidence are the following:

- 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.
- The method does not miss any Pareto optimal solution that is more preferred (with a given tolerance) for the DM than the one chosen.
- 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.

The property "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" was again operationalised to four items:

- After this iteration, I know more about the problem.
- I think that the solution I found is the best one.
- I obtained a clear idea of the values that the objectives can simultaneously achieve.
- I obtained a clear idea of the possible choices available similar to the solutions I was interested in.

It is important to find out if the DM has learned enough about the offered solutions, because in order to feel confident about their final choice, the DM has to understand the problem well enough (Afsar et al., 2024; Hakanen et al., 2022). The items above give us a sense of whether the DM has learned about the objectives, trade-offs, conflict degrees, and how the MPS has been generated. If the DM cannot identify the MPS among the available solutions, it is unclear whether they have understood the problem or the method well enough.

The property "The method does not miss any Pareto optimal solution that is more preferred (with a given tolerance) for the DM than the one chosen" has been measured with one Likert-scale item, which is also open-ended. It should be considered if the property can be operationalised to only one Likert-scale item. The purpose of this property in measuring the DM's satisfaction and confidence is to find out if the method has successfully captured the preferences of the DM and offered them solution that reflect them. The DM's confidence in the final solution requires that the method has presented the DM enough solutions that match their preconditions and preferences.

The third property "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" is operationalised to four following items:

- I am satisfied with the final solution.
- Overall, I am satisfied with the ease of completing this task.
- Overall, I am satisfied with the amount of time it took to complete this task.
- Overall, I am satisfied with the support information (online help, messages, documentation) when completing this task.

The last three items come from ASQ (after-scenario questionnaire) (Lewis, 1995), a validated measurement scale to measure user satisfaction. It was originally developed for the purpose of usability testing, but its general nature allows it to be applied to different research contexts in human-computer interaction (Afsar et al., 2024; Lewis, 1995).

This property aims to evaluate how confident the DM is that the solution process generated the best possible solution and that they were able to recognise and choose it. The property evaluates both the DM's feeling of confidence and the ability of the interaction process to promote such a feeling in the DM. The first item assesses the DM's satisfaction with the final solution, and the other three measure it as well. But even though the three items from ASQ measure the satisfaction and the feeling itself is mentioned in the items, the context of feeling of satisfaction is in the experienced task difficulty, time spent completing the task, and the amount of help participants got during the process. In other words, the items reflect the metacognitive feeling of difficulty the DM experienced during the solution process.

Metacognitive experiences, such as metacognitive feelings and judgments, appear in learning situations, like problem solving (Efklides, 2006). Feeling of confidence and feeling of satisfaction monitor the outcome of cognitive processing

(Efklides, 2006). Feeling of difficulty monitors the fluency of cognitive processing, and it consists of multiple factors, including objective task difficulty, individual cognitive abilities, emotions, and the affective tone of instructions, to name a few (Efklides, 2006; Efklides, 2002). According to Nelson (1996), confidence is a retrospective metacognitive monitoring component, that indicates how sure the person is that the retrieved answer is correct. Efklides (2002), however, implies that confidence monitors how the person reached the answer (was the process fluent or with interruptions). Considering these two perspectives, the confidence is a feeling a person experiences, but it reveals information about the process also.

Feeling of confidence is related to feeling of difficulty (Efklides, 2002). In their study from 1996, Efklides made an observation that retrospective feeling of difficulty is negatively related to feeling of confidence (Efklides, 2002). According to Efklides (2002), this is in accordance with previous findings made by Robinson et al. (1997), that confidence is associated with cognitive effort and time spent on the task. Based on these findings, the items borrowed from ASQ can be claimed to measure the DM's experience of the learning process and reflect the feeling of difficulty of the task in the context of interactive multiobjective optimisation. And since the feeling of confidence towards the final answer or solution is influenced by the feeling of difficulty (Efklides, 2002), these items are suitable for assessing the DM's confidence in the final solution.

However, to measure the confidence reliably, the first questionnaire item regarding the satisfaction with the final solution would be advisable to be left out of scale development. There is interrelation between metacognitive experiences, and the feeling of satisfaction and feeling of confidence do correlate as well (Efklides, 2002). Their connection is contrary than what the operationalisation here assumes: the feeling of satisfaction is influenced by the feeling of confidence, not the other way around.

4.4 Research question

There are no known validated scales to assess desirable properties of interactive multiobjective optimisation methods (Afsar et al., 2024). The purpose of this research is to investigate the applicability of the results from the previous study by Afsar et al. (2024) to form a reliable scale or a set of scales. The following research question guides this study:

> **RQ1**: Is the developed scale or a set of scales a reliable measure for assessing interactive multiobjective optimisation methods?

4.5 Research data

This study utilises already gathered research data from Afsar et al. (2024). Participants (N = 164) were mathematics students from the Faculty of Economics and Business Studies of the University of Malaga. Multiobjective optimisation was a familiar field for all of them. The participants were divided into three groups with one interactive multiobjective optimisation method assigned to each group: E-NAUTILUS (n =64), NIMBUS (n =44), and RPM (n = 56) (Afsar et al., 2024).

The participants were introduced to the multiobjective optimisation problem to be solved and the user interface of each method (Afsar et al., 2024). Participants were also given supplementary documentation consisting of detailed information about the problem and the interactive method allocated to them, so they could carefully consider their preferences (Afsar et al., 2024).

A questionnaire included thirty items, from which 25 items are considered in this study. Items were both open-ended and graded on a given scale (either Likert-scale or semantical differential) (Afsar et al., 2024). Four items were asked during the solution process and the remaining 26 after the solution process had ended. Some items were asked twice between different iterations. (Afsar et al., 2024).

4.6 Data analysis

Factor analyses are statistical procedures that are used to identify the number of distinct constructs assessed by a set of measures (Fabrigar & Wegener, 2011). Fabrigar and Wegener (2011) explain the factor analysis as follows: “Factor analysis is used as a means of arriving at a more parsimonious representation of the underlying structure of correlations among a set of measured variables”. Factor analysis is not the only procedure to examine the structure of correlations among certain variables: its use is beneficial in a case where researchers want to know how many constructs a set of measured variables is assessing and what they might be, but investigating the causal relation of said constructs is not yet relevant (Fabrigar & Wegener, 2011). Factor analysis is broadly used to develop measurement instruments (Fabrigar & Wegener, 2011).

According to Fabrigar and Wegener (2011), first a researcher has to decide between an exploratory or a confirmatory factor analysis. An exploratory factor analysis is suitable for situations where there are no expectations about the number of factors, and which variables will be influenced by the same factors (Fabrigar & Wegener, 2011). For the purposes of this study, the exploratory factor analysis is more suitable, because there are no strong preliminary expectations for the factor solution. The next decision to make is whether to conduct the analysis based on the common factor model or PCA (Fabrigar & Wegener, 2011). The main difference between them is that common factors are unobservable latent variables (Fabrigar & Wegener, 2011), and if the researchers do not assume latent variables to exist in data or do not assume that they are to be found, PCA is more suitable. PCA was chosen to conduct the research data analysis, because it was not assumed beforehand that unobservable, hidden variable would explain the phenomenon behind the data.

There exists multiple recommendations for determining an acceptable sample size to conduct factor analysis (MacCallum et al., 1999; Mundfrom et al., 2005). An absolute minimum sample size of 100 has been suggested (Gorsuch, 1983), and some authors recommend a minimum sample size of 200 to 500 (MacCallum et al., 1999; Comrey & Lee, 1992). For Comrey and Lee (1992), the sample size of 100 is poor. However, most authors would appear to accept a minimum sample size of 200 and 5-to-1 participant-to-variable ratio, whichever is greater (Howard, 2016). If the suggestion to apply the 5-to-1 participant-to-variable ratio is accepted, the research data utilised in this study is adequate for performing the PCA.

The results of factor analysis do not offer simple answers, but useful information that can be interpreted to understand the underlying phenomena (Clark & Watson, 1995). Simply following few rules is not optimal (Clark & Watson, 1995). When performing factor analysis or PCA, there are multiple choices to be made regarding extraction, rotation, and number of factors, for instance (O'Connor, 2000). The most crucial decision, however, is selecting the number of factors or components to retain (Zwick & Velicer, 1986; Glorfeld, 1995; O'Connor, 2000). One way to do that is to examine the scree plots of eigenvalues, which is available on the SPSS software (O'Connor, 2000). The problem in eigenvalues-greater-than-one rule is that it can over- or underestimate the number of components, and generate unreliable results (O'Connor, 2000; Zwick & Velicer, 1986).

Another method to determine the number of components to retain is *parallel analysis*, which extracts eigenvalues from random data sets that correspond to the number of cases and variables in the actual data set (O'Connor, 2000). The method generates a set of random data matrices that have the same number of observations for each variable than in the actual data set (O'Connor, 2000). Eigenvalues for the correlation matrices of the original data set as well as the correlation matrices derived from random data sets are calculated (O'Connor, 2000). After that, the method compares the eigenvalues of the actual data to the eigenvalues extracted from the random data (O'Connor, 2000). According to O'Connor (2000), it is recommended to use the eigenvalues that correspond to the desired percentile (typically the 95th) of the distribution of random data eigenvalues as a comparison baseline.

If the parallel analysis is run by using the desired percentile of the distributions of random data eigenvalues, the number of random data sets generated should be carefully chosen (O'Connor, 2000). According to O'Connor (2000), if the chosen percentile is multiplied by the number of data sets and then divided by 100, the result should be an integer. Factors or components are retained as long as the eigenvalue from the actual (raw) data is greater than the eigenvalue from the random data (O'Connor, 2000). The SPSS commands (syntax) for parallel analysis appears in Appendix C in O'Connor's (2000) paper. The user defines the number of cases, variables, data sets, and the desired percentile for the analysis before running it (O'Connor, 2000).

5 RESULTS

This chapter presents the results of statistical data analysis. Based on the reasoning of operationalisations in Section 4.3., the item "I am satisfied with the final solution", was left out of the analysis.

5.1 Parallel analysis

Before performing the PCA, parallel analysis using the O'Connor syntax (2000) was first conducted to determine the number of components to retain. The parallel analysis was made in SPSS version 28.0.1.1. First, the syntax was checked and corrected to perform the analysis in a desired way. The method (compute kind) was chosen to be the PCA, and compute randtype was assigned 1, meaning normally distributed random data generation parallel analysis. These and other syntax choices are presented in Table 3.

TABLE 3 Choices in O'Connor (2000) syntax

compute ndatsets	1000
compute percent	95
compute kind	*1* (principal component analysis)
compute randtype	*1* (normally distributed random data generation parallel analysis)

The results of parallel analysis were interpreted against the 95th percentile. The value of raw data should be higher than the value of the 95th percentile eigenvalue. Components are retained as long as the eigenvalue from the raw data is greater than the eigenvalue from the random data (O'Connor, 2000). The parallel analysis suggests that there are three components to retain from these questionnaire items. The values of the components are presented in Table 4.

TABLE 4 Raw data and 95th percentile random data Eigenvalues

Raw Data	95th Percentile
7.702	1.939
2.536	1.748
1.864	1.632
1.432	1.547

As Table 4 shows, the value of the raw data in the fourth component is less than the value of the 95th percentile. Therefore, the PCA should retain three components.

5.2 PCA

PCA was conducted using the Principal components method with Promax-rotation. Promax-rotation was chosen because the components are assumed to correlate with each other. Three components were assigned to be retained from the analysis. Bartlett's test of sphericity ($\chi 2(276) = 1658.138$, $p <. 001$) was statistically significant. Kaiser Meyer-Olkin test of sampling adequacy (KMO) was .865. Therefore, the research data can be used to conduct a factor analysis.

After the first iteration, four items had a communality less than .40. These items were the following:

- The preference information was easy to provide. (.153)
- The solution(s) I obtained reflects my preference information well. (.295)
- After this iteration, I know more about the problem. (.308)
- I was able to return to previous solutions whenever I needed in the solution process. (.242)

These items were iteratively removed, until all the items had a communality greater than or equal to .40. In the fifth and sixth iteration, two cross-loaded items in the pattern matrix were removed, one at each iteration. The items were "I felt tired" (the factor loading for component 1 was - .586 and component 3 .506) and "I was able to reflect my actual preferences when providing the information required by the method" (the factor loading for component 1 was .412 and component 2 .427). After the seventh iteration, the pattern matrix included only items that had a communality greater than or equal to .40 and there were no cross-loadings. At this point, there were three components. The first component had nine items, the second component had five items and the third component had four items. The third component was left as it was and named Cognitive load.

The first component and the second component did not describe the concepts of decision-making and satisfaction coherently, so they were modified to describe these concepts better. The second component (later named as "Satisfaction") included five items at this point. The following two items were removed from the component:

- In general, the method reacted to the preference information I provided.
- I felt I was in control during the solution process.

The item "In general, the method reacted to the preference information I provided" had a factor loading of .584, and the item "I felt I was in control during the solution process" had a factor loading of .505. These factor loadings were weaker compared with other items in this component. After removing the aforementioned two items, the component was named Satisfaction. It describes the satisfaction towards the result of the solution process and the DM's own performance.

Before modifying the first component to better describe the concept of decision-making in the context of interactive multiobjective optimisation, the component had nine items. They reflected very different parts of the user's experience, including easiness to learn, feeling of satisfaction, obtaining the clear idea of different solutions presented in the method, and the feeling of comfort. There was also an item that referred to the method's functionalities ("The method has all the necessary functionalities"). The following six items were removed from the component:

- It was easy to learn to use this method.
- Overall, I am satisfied with the support information when completing this task.
- Overall, I am satisfied with the amount of time it took to complete this task.
- Overall, I am satisfied with the ease of completing this task.
- I felt comfortable using this interactive method.
- The method has all the necessary functionalities.

The items were removed, because they did not describe the characteristics of the decision-making specific to the multiobjective optimisation. The following three items were selected to the final component:

- I obtained a clear idea of the values that the objectives can simultaneously achieve.
- I obtained a clear idea of the possible choices available similar to the solutions I was interested in.
- It was easy to explore solutions with different conflicting values of the objective functions.

These three items describe the characteristics of the decision-making specific to the multiobjective optimisation, and the method's ability to support the DM to make a decision. Out of the nine items, these three items had the lowest factor loadings (.617, .542, .605). The item "The method has all the necessary functionalities" also had one of the lowest loadings (.580), but it was removed, because it did not adequately describe the degree to which the method supports the DM to make a decision. The component was named Decision-making support. The final components are presented in Table 5.

TABLE 5 Final components

Items	**Satisfaction**	**Cognitive load**	**Decision-making support**
I think that the solution I found is the best one.	.914		
I am satisfied with the solution I chose.	.885		
I am satisfied with my performance in finding the final solution.	.790		
I had to work hard to find the final solution.		.796	
A lot of mental activity was required (e.g., thinking, deciding, and remembering).		.764	
I felt frustrated in the solution process (e.g., insecure, discouraged, irritated, stressed).		.716	
I took too many iterations to arrive to the acceptable solution.		.694	
I obtained a clear idea of the values that the objectives can simultaneously achieve.			.863
I obtained a clear idea of the possible choices available similar to the solutions I was interested in.			.781
It was easy to explore solutions with different conflicting values of the objective functions.			.775

Satisfaction-component correlates very weakly with the Cognitive load -component (r = - .088) and moderately with Decision-making support -component (r = .505). Cognitive load -component has also a very weak correlation with the Decision-making support -component (r = - .064). The items of each component were calculated into a sum variable, after which the internal consistency of each variable was measured. Cronbach's alphas are at an acceptable level. Satisfaction-variable has the highest value of alpha (α = .852), while Cognitive load (α = .730) and Decision-making support (α = .738) score lower levels of alpha value. Because the correlations between the components are weak, they cannot be combined into a single scale. They form a set of individual scales, that measure the DM's perceived cognitive load and satisfaction, and the degree of the decision-making support that the method provides.

6 DISCUSSION

This chapter reviews the results of statistical data analysis and discusses their meaning, while proposing suggestions for future research. The research goal of this thesis was to examine if the research data collected by Afsar et al. (2024) would be suitable to construct a reliable scale or scales to assess interactive multiobjective optimisation methods. Reliability and validity of the developed scales are assessed, and theoretical and practical contributions of these results are reviewed. Limitations of the study are also discussed.

PCA resulted in three components, which are named Satisfaction, Cognitive load, and Decision-making support. The components were calculated into sum variables. Cronbach alphas for each of the sum variables are at an acceptable level, which means that the components are internally coherent. In other words, questionnaire items in each component measure the same construct reliably. Correlations between the components are relatively weak, except the correlation between Satisfaction-component and Decision-making support -component. Correlations indicate that these three components measure distinct constructs of interaction between the DM and the interactive multiobjective optimisation methods and have a poor capability to be combined into a single scale. They form a set of individual scales (*Satisfaction*, *Cognitive load*, and *Decision-making support*) to assess interactive multiobjective optimisation methods. Future research, where these scales would be tested with different interactive methods and optimisation problems, is needed to examine their validity and justify their use in developing and testing interactive methods.

Originally, in Afsar et al. (2023) and Afsar et al. (2024), the questionnaire items were connected to specific desirable properties (see Table 2). When conducting PCA, it was not assumed that this structure would remain in retained components. The developed scales measure aspects of decision-making that are also identified in desirable properties and that characterise the performance of interactive methods, like low cognitive burden and the method's support for learning. The developed scales are not, however, supposed to measure the desirable properties of interactive multiobjective optimisation methods per se, but rather the aspects of decision-making present in interaction between the DM and the interactive method.

The Satisfaction-component measures the DM's satisfaction towards the final result of multiobjective optimisation, and their own performance in finding that. The item "I am satisfied with my performance in finding the final solution" is connected to cognitive load in Afsar et al. (2023; 2024) and it is part of NASA-TLX, the standardised questionnaire often applied to measure user cognitive load in HCI research. The internal consistency of the component is however satisfactory, which indicates that the items measure a same construct. In HCI, satisfaction is often understood as part of usability. When performing usability evaluations, satisfaction has been operationalised to perception of outcomes, among other things. *Perception of outcomes* includes assessments of user's own performance as well as the task outcome (Hornbæk, 2006). Following this research line, the DM's satisfaction towards their own performance can be evaluated together with the satisfaction towards the final solution. If metacognitive experiences are considered, the interconnections between feeling of satisfaction, feeling of confidence, and feeling of difficulty, and their impact on decision-making is a little more complicated (cf. Efklides, 2001; 2002; 2006), and exploring their effect on decision-making in the context of interactive multiobjective optimisation methods could be beneficial.

The Cognitive load -component includes items from NASA-TLX, and one item developed by Afsar et al. (2023, 2024): "It took too many iterations to arrive to the acceptable solution". Regardless of the satisfactory level of internal consistency, including this item to measure the DM's cognitive load during the use of interactive method, is not uncomplicated. In previous studies (Afsar et al., 2021; 2023; 2024) and in this thesis, it is assumed that the excess number of iterations negatively affect the DM's cognitive load. This may be true, but the number of iterations does not necessarily impact the cognitive load directly. In Afsar et al. (2024), participants spent more time and iterations with E-NAUTILUS but reported being less tired. The number of iterations was also deemed appropriate (Afsar et al., 2024). This indicates that the time to be spent reaching the MPS is necessarily not associated with increase in experiencing cognitive load. In Afsar et al. (2023), NIMBUS was cognitively more demanding, but it was the method that most participants said they would use again. According to the participants, NIMBUS reacted better to the DM's preference information (Afsar et al., 2023). This allowed participants learning more about the problem and reaching a more satisfactory solution (Afsar et al., 2023). The RPM was rated easier and simpler to use (Afsar et al., 2023). Afsar et al. (2024) reported similar results about RPM being the least liked method, even though its apparent simplicity. The reason may be that in order to learn, appropriate levels of germane cognitive load is needed to construct schemas (see Section 3.2).

In light of the knowledge regarding the DMs' experiences with interactive multiobjective optimisation methods and the concept of cognitive load, it is not surprising that Cognitive load -component correlates so poorly with Satisfaction and Decision-making support -components. Based on the results reported by

(Afsar et al., 2024), participants are willing to cope with higher cognitive load levels in order to learn more about the problem and attain better results. In the context of interactive multiobjective optimisation, cognitive load seems to have unique features and decreasing cognitive load may not be beneficial in all cases. Multiobjective optimisation problems are known to be complex, and the DM's motivation to engage in high-demanding cognitive tasks might be stronger than in everyday tasks. Considering all of this, it would be advisable to develop an entirely novel measurement instrument to measure cognitive load in decision-making with interactive multiobjective optimisation methods. The scale developed in this thesis might be valid to measure cognitive load, but there is a possibility that it misses important aspects of human decision-making and problem solving in this context. Either way, future research is needed.

The Decision-making support -component includes three items from different desirable properties. The item "It was easy to explore solutions with different conflicting values of the objective functions" has been originally connected to the desirable property which describes how well the method reacts to changes in DM's preference information during the solution process. The other two items ("I obtained a clear idea of the values that the objectives can simultaneously achieve" and "I obtained a clear idea of the possible choices available similar to the solutions I was interested in") are connected to the desirable property which evaluates how the method supports the DM's learning process. The latter desirable property is part of the satisfaction and confidence -research question in Afsar et al. (2024), and the former of capturing preferences and responsiveness -research question. The Decision-making support -component is considered to measure the method's ability to support the decision-making process of the DM. The items in this component reflect the unique features of decision-making process in interactive multiobjective optimisation methods. The DM guides the solution process by providing preference information, and learning about the trade-offs and different possible solutions is at the core of finding the MPS. The correlation between Satisfaction-component and Decision-making support -component is medium. This is expected, because in order to be able to identify the MPS, the DM needs to learn about the problem.

The RPM method is based on the Simon's theory of satisficing organism, and the NAUTILUS method is designed to diminish the effect of anchoring and adjustment biases as discussed in Section 3.1. However, published studies about interactive methods do not usually mention that a specific theory regarding human decision-making and judgment would have been utilised in the development of the method. The idea of finding the satisfactory solution instead of the absolute best one is an important aspect of multiobjective optimisation. Yet, the RPM method that is based on this idea was the least liked method in Afsar et al. (2024). E-NAUTILUS has been popular among the participants (Afsar et al., 2023; 2024), but it has not been examined whether it has any connection to the prevalence of cognitive biases. If the notions of FTT are examined, research tradition regarding heuristics and biases approach might not be the most useful to consider. Expert decision-making seems more relevant in the context of interactive multiobjective optimisation, but it cannot be known for sure without studying DMs more.

In order to determine which theories of human decision-making and judgment are the most useful ones to explain the decision-making behaviour of DMs, the specific qualities of DMs who solve optimisation problems by using interactive methods should be examined. This requires both qualitative and quantitative analysis of the decision-making and learning processes typical for DMs in the context of interactive multiobjective optimisation. In future research, the focus should be on how DMs learn and how cognitive load affects learning (in this context), what affects DMs' satisfaction in the final solution and how, are feeling of difficulty and feeling of confidence relevant in determining DMs' satisfaction, does being an expert regarding the multiobjective optimisation problem domain affect the decision-making of DMs, and do cognitive biases affect the decision-making processes of DMs. This kind of research has not yet been done to our knowledge. The most prominent theories of decision-making and judgment with interactive multiobjective optimisation methods are presented in Chapter 3. Chapter 3 also introduces the concepts of usability, satisfaction and cognitive load in the HCI. The literature presented in Chapter 3 is important to consider in future research. Other research from the fields of cognitive science, HCI, and psychology also offer valuable knowledge regarding the qualities of the DM.

The study reported in this thesis has a few limitations. The sample size is sufficient to perform PCA, but bigger sample sizes would offer more reliable results. The participants in the study by Afsar et al. (2024) were mathematics students, which limits the generalisability of the results. The research data used in this study was not originally intended for the purposes of scale development, so the operationalisation process might not have been paid enough attention to when designing the questionnaire items in Afsar et al. (2024). Considering the future research on interactive multiobjective optimisation methods in the field of HCI, it would be advisable to approach the subject by operationalising the questionnaire items without the desirable properties. When developing a scale, it is important to operationalise the questionnaire items carefully. If the operationalisations are drawn from the desirable properties, important aspects of psychology of decision-making can be overlooked. Desirable properties describe the preferable qualities of interactive methods, and they are useful in evaluating the performance of interactive methods (Afsar et al., 2021). But in order to develop these methods to better consider the human decision-making processes and the elements of interaction between the DM and the method, knowledge of DMs' properties should guide the research process and questionnaire item operationalisations, especially in scale development.

7 CONCLUSIONS

The objective of this thesis was to examine if the research data gathered in the previous study (Afsar et al., 2024) can be utilised to form a reliable measurement scale or scales to assess interactive multiobjective optimisation methods. The research question that guided the study in this thesis was *"Is the developed scale or a set of scales a reliable measure for assessing interactive multiobjective optimisation methods?"*. To understand the concepts of multiobjective optimisation, interactive multiobjective optimisation methods, human decision-making, and interaction between the DM and the interactive methods, key concepts were defined. These concepts are utilised to evaluate the validity of the developed scales.

Following the definitions of the key concepts, a quantitative data analysis was performed in order to construct a scale or scales. The analysed research data (N = 164) had been collected by a between-subjects experiment, where each participant used only one method. The analysis included 25 questionnaire items from the study. The results of parallel analysis and PCA indicated that three components are to be found from the data. These components were named Satisfaction, Cognitive load, and Decision-making support. Each component was calculated into a sum variable and their internal consistency was evaluated using Cronbach's alpha. The alpha values were at an acceptable level. The correlations between the components were low, which indicates that they measure separate constructs of the interaction between the DM and the interactive multiobjective optimisation methods. These components form a set of individual measurement scales (*Satisfaction*, *Cognitive load*, *Decision-making support*), that can reliably be used to assess interactive multiobjective optimisation methods. Further evaluations of their validity require more research.

This has been the first attempt to form a scale that reliably assesses interactive multiobjective optimisation methods. Useful knowledge from human decision-making and judgment as well as HCI has been provided to identify factors affecting the use of interactive multiobjective optimisation methods. The definitions of the key concepts and the results presented in this thesis offer future research suggestions. First of all, qualitative analysis of the decision-making processes of DMs is needed to identify which theories of human decision-making and judgment are relevant to consider in future research. In addition, deeper understanding and careful definition of cognitive load is an important topic to consider. Examining the relations between cognitive load, feeling of satisfaction, learning, and decision-making in the context of interactive multiobjective optimisation is needed. In future studies it is also important to carefully operationalise the measured constructs.

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.

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.

Afsar, B., Silvennoinen, J., Ruiz, F., Ruiz, A. B., Misitano, G., & Miettinen, K. (2024). An experimental design for comparing interactive methods based on their desirable properties. *Annals of Operations Research*, 1-22.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. *Psychology of Learning and Motivation*, *2*, 89-195.

Baddeley, A. (1992). Working memory. *Science, 255*(5044), 556–559

Bargas-Avila, J. A., & Hornbæk, K. (2011). Old wine in new bottles or novel challenges: a critical analysis of empirical studies of user experience. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2689-2698). Association for Computing Machinery.

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. Słowiński (Eds.), *Multiobjective optimization: Interactive and Evolutionary Approaches* (pp. 405-433). Springer Berlin Heidelberg.

Bevan, N. (2010). Extending the concept of satisfaction in ISO standards. In *Proceedings of the KEER 2010 International Conference on Kansei Engineering and Emotion Research*.

Bevan, N., Carter, J., & Harker, S. (2015). ISO 9241-11 revised: What have we learnt about usability since 1998?. In *Proceedings of the 17th International Conference on Human-Computer Interaction: Design and Evaluation* (pp. 143-151). Springer.

Borragán, G., Slama, H., Bartolomei, M., & Peigneux, P. (2017). Cognitive fatigue: A time-based resource-sharing account. *Cortex*, *89*, 71-84.

Branke, J., Deb, K., Miettinen, K., & Słowiński, R. (Eds). (2008). *Multiobjective Optimization: Interactive and Evolutionary Approaches*. Springer Berlin Heidelberg.

Braver, T. S. (2012). The variable nature of cognitive control: A dual mechanisms framework. *Trends in Cognitive Sciences, 16*(2), 106-113.

Brainerd, C., & Reyna, V. (1990). Gist is the grist: Fuzzy-trace theory and the new intuitionism. *Developmental Review, 10*(1), 3-47.

Brainerd, C. J., & Reyna, V. F. (2001). Fuzzy-trace theory: dual processes in memory, reasoning, and cognitive neuroscience. *Advances in Child Development and Behavior*, *28*, 41-100.

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., & Corner, J. (1997). The effects of anchoring in interactive MCDM solution methods. *Computers & Operations Research, 24*(10), 907-918.

Buchanan, J., & Gardiner, L. (2003). A comparison of two reference point methods in multiple objective mathematical programming. *European Journal of Operational Research*, *149*(1), 17-34.

Carroll, J. M. (1997). Human-computer interaction: Psychology as a science of design. *Annual Review of Psychology, 48*(1), 61-83.

Chin, J. P., Diehl, V. A., & Norman, K. L. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 213-218). Association for Computing Machinery.

Clark, L. A., & Watson, D. (1995). Constructing Validity: Basic Issues in Objective Scale Development. *Psychological Assessment, 7*(3), 309-319.

Comrey, A. L. & Lee H. B. (1992). *A first course in factor analysis*. Hillsdale, NJ: Erlbaum.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334.

Efklides, A. (2001). Metacognitive experiences in problem solving: Metacognition, motivation, and self-regulation. In A. Efklides, J. Kuhl, & R. M. Sorrentino (Eds.), *Trends and Prospects in Motivation Research* (pp. 297-323). Kluwer Academic Publishers.

Efklides, A. (2002). The systemic nature of metacognitive experiences: Feelings, judgments, and their interrelations. In P. Chambres, M. Izaute & PJ. Marescaux (Eds.), *Metacognition: Process, Function and Use* (pp. 19-34). Springer Boston.

Efklides, A. (2006). Metacognition and affect: What can metacognitive experiences tell us about the learning process?. *Educational Research Review*, *1*(1), 3-14.

Evans, J. S. B. T. (1984). Heuristic and analytic processes in reasoning. *British Journal of Psychology*, *75*, 451-468

Evans, J. S. B. (1989). *Bias in human reasoning: Causes and consequences*. Erlbaum.

Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, *8*(3), 223-241.

Epstein, S. (1994). Integration of the cognitive and the psychodynamic unconscious. *American Psychologist*, *49*(8), 709.

Eskelinen, P., Miettinen, K., Klamroth, K., & Hakanen, J. (2010). Pareto navigator for interactive nonlinear multiobjective optimization. *OR Spectrum*, *32*, 211-227.

Fabrigar, L. R., & Wegener, D. T. (2011). *Exploratory factor analysis*. Oxford University Press.

Frisch, D., & Clemen, R. T. (1994). Beyond expected utility: rethinking behavioral decision research. *Psychological Bulletin*, *116*(1), 46.

Gigerenzer, G. (2018). The bias bias in behavioral economics. *Review of Behavioral Economics*, *5*(3-4), 303-336.

Gigerenzer, G. (2020). What is bounded rationality? In R. Viale (Ed.), *Routledge Handbook of Bounded Rationality* (pp. 55-69). Routledge.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic Decision Making. *Annual Review of Psychology, 62*(1), 451-482.

Gilovich, T., & Griffin, D. (2002). Introduction - Heuristics and biases: Then and now. In T. Gilovich, D. Griffin & D. Kahneman (Eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment* (pp. 1-18). Cambridge university press.

Glorfeld, L. W. (1995). An improvement on Horn's parallel analysis methodology for selecting the correct number of factors to retain. *Educational and Psychological Measurement*, *55*(3), 377-393.

Gorsuch, R. L. (1983). *Factor analysis* (2. ed.). Hillsdale, NJ: Erlbaum.

Hakanen, J., Radoš, S., Misitano, G., Saini, B. S., Miettinen, K., & Matković, K. (2022). Interactivized : Visual Interaction for Better Decisions with Interactive Multiobjective Optimization. *IEEE Access*, *10*, 33661-33678.

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

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.

Hartson, H. R. (1998). Human–computer interaction: Interdisciplinary roots and trends. *Journal of Systems and Software*, *43*(2), 103-118.

Hassenzahl, M. (2003). The thing and I: Understanding the relationship between user and product. In M. Blythe, C. Overbeeke, A. F. Monk, & P. C. Wright (Eds.), *Funology: From usability to enjoyment* (pp. 31–42). Dordrecht: Kluwer.

Hastie, R. (2001). Problems for judgment and decision making. *Annual Review of Psychology*, *52*(1), 653-683.

Haugeland, J. (1978). The nature and plausibility of cognitivism. *Behavioral and Brain Sciences, 1*(02), 215–226

Helm, R. K., McCormick, M. J., & Reyna, V. F. (2017). Expert decision making: A fuzzy-trace theory perspective. In L. J. Ball & V. A. Thompson (Eds.), *International Handbook of Thinking and Reasoning* (pp. 289-303). Routledge.

Hollender, N., Hofmann, C., Deneke, M., & Schmitz, B. (2010). Integrating cognitive load theory and concepts of human–computer interaction. *Computers in Human Behavior*, *26*(6), 1278-1288.

Hornbæk, K. (2006). Current practice in measuring usability: Challenges to usability studies and research. *International Journal of Human-Computer Studies*, *64*(2), 79-102.

Howard, M. C. (2016). A Review of Exploratory Factor Analysis Decisions and Overview of Current Practices: What We Are Doing and How Can We Improve? *International Journal of Human-Computer Interaction, 32*(1), 51-62.

Hwang, C.–L. & Masud, A. S. M. (1979). Multiple Objective Decision Making - Methods and Applications: A State-of-the-Art Survey. In M. Beckmann & H. P. Künzi (Eds.), *Lecture Notes in Economics and Mathematical Systems*. Springer-Verlag Berlin Heidelberg.

International Organization for Standardization. (2018). *Ergonomics of human-system interaction – Part 11: Usability: Definitions and concepts* (ISO Standard No. 9241-11:2018). https://www.iso.org/standard/63500.html

Jimenez, C., Lozada, P., & Rosas, P. (2016). Usability heuristics: A systematic review. In *2016 IEEE 11th Colombian Computing Conference (CCC)* (pp. 1-8). IEEE.

Jokinen, J. P. (2015). User psychology of emotional user experience. *Jyväskylä studies in computing*, (213).

Jones, B. D. (1999). Bounded rationality. *Annual Review of Political Science*, *2*(1), 297-321.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), 430-454.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*(4), 237.

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, *47*(2), 263–291.

Kirakowski, J., & Dillon, A. (1988). The computer user satisfaction inventory (CUSI): Manual and scoring key. *Ireland: University College of Cork: Human Factors Research Group*.

Kosch, T., Karolus, J., Zagermann, J., Reiterer, H., Schmidt, A., & Woźniak, P. W. (2023). A Survey on Measuring Cognitive Workload in Human-Computer Interaction. *ACM Computing Surveys, 55*(13), 1-39.

Lallemand, C., Gronier, G., & Koenig, V. (2015). User experience: A concept without consensus? Exploring practitioners' perspectives through an international survey. *Computers in Human Behavior*, *43*, 35-48.

LaLomia, M. J., & Sidowski, J. B. (1990). Measurements of computer satisfaction, literacy, and aptitudes: A review. *International Journal of Human-Computer Interaction*, *2*(3), 231-253.

Law, E. L. C., Roto, V., Hassenzahl, M., Vermeeren, A. P., & Kort, J. (2009). Understanding, scoping and defining user experience: a survey approach. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 719-728). Association for Computing Machinery

Lewis, J. R. (1992). Psychometric evaluation of the post-study system usability questionnaire: The PSSUQ. In *Proceedings of The Human Factors Society Annual Meeting* (pp. 1259-1260). Sage Publications.

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.

Lewis, J. R. (2014). Usability: Lessons Learned ... and Yet to Be Learned. *International Journal of Human-Computer Interaction, 30*(9), 663-684.

Litwin, M. S., & Fink, A. (1995). *How to measure survey reliability and validity* (Vol. 7). Sage.

Luque, M., Miettinen, K., Eskelinen, P., & Ruiz, F. (2009). Incorporating preference information in interactive reference point methods for multiobjective optimization. *Omega*, *37*(2), 450-462.

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

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4*(1), 84–99.

Martin, S. (2014). Measuring cognitive load and cognition: metrics for technology-enhanced learning. *Educational Research and Evaluation*, *20*(7-8), 592-621.

Mellers, B. A., Schwartz, A., & Cooke, A. D. J. (1998). Judgment and decision making. *Annual Review of Psychology*, *49*(1), 447-477.

Miettinen, K. (1999). *Nonlinear multiobjective optimization*. Kluwer Academic Publishers.

Miettinen, K. (2008). Introduction to Multiobjective Optimization: Noninteractive Approaches. In J. Branke, K. Deb, K. Miettinen & R. Słowiński (Eds.), *Multiobjective Optimization: Interactive and Evolutionary approaches* (pp. 1-26). Springer Berlin Heidelberg.

Miettinen, K., Eskelinen, P., Ruiz, F., & Luque, M. (2010). NAUTILUS method: An interactive technique in multiobjective optimization based on the nadir point. *European Journal of Operational Research*, *206*(2), 426-434.

Miettinen, K., Ruiz, F., & Wierzbicki, A. P. (2008). Introduction to multiobjective optimization: interactive approaches. In *Multiobjective optimization: interactive and evolutionary approaches* (pp. 27-57). Springer Berlin Heidelberg.

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

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.

Miyake, S. (2001). Multivariate workload evaluation combining physiological and subjective measures. *International Journal of Psychophysiology*, *40*(3), 233-238.

Mundfrom, D. J., Shaw, D. G., & Ke, T. L. (2005). Minimum Sample Size Recommendations for Conducting Factor Analyses. *International Journal of Testing*, *5*(2), 159–168

Nelson, T. O. (1996). Consciousness and metacognition. *American Psychologist*, *51*(2), 102.

Nielsen, J. (2005). *Ten usability heuristics.* https://pdfs.semanticscholar.org/5f03/b251093aee730ab9772db2e1a8a7eb8522cb.pdf.

Niiniluoto, I. (1980). *Johdatus tieteenfilosofiaan: Käsitteen- ja teorianmuodostus*. Otava.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory* (3. ed.). McGraw-Hill.

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments, & Computers*, *32*(3), 396-402.

Olson, G. M., & Olson, J. S. (2003). Human-computer interaction: Psychological aspects of the human use of computing. *Annual Review of Psychology*, *54*(1), 491-516.

Ojalehto, V. & Miettinen, K. (2019). DESDEO: An open framework for interactive multiobjective optimization. In S. Huber, M. Geiger & A. de Almeida (Eds.), *Multiple Criteria Decision Making and Aiding* (pp. 67-94). Springer Cham.

Osman, M. (2004). An evaluation of dual-process theories of reasoning. *Psychonomic Bulletin & Review*, *11*(6), 988-1010.

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist*, *38*(1), 1-4.

Quiñones, D., & Rusu, C. (2017). How to develop usability heuristics: A systematic literature review. *Computer Standards & Interfaces*, *53*, 89-122.

Reyna, V. F. (2005). Fuzzy-Trace Theory, Judgment, and Decision-Making: A Dual-Processes Approach. In C. Izawa & N. Ohta (Eds.), *Human Learning and Memory: Advances in Theory and Application: The 4th Tsukuba International Conference on Memory* (pp. 239–256). Lawrence Erlbaum Associates Publishers.

Reyna, V. F. (2013). Intuition, reasoning and development: A fuzzy-trace theory approach. In P. G. Barrouillet, P. Barrouillet, & C. Gauffroy (Eds.), *The development of thinking and reasoning* (pp. 193-220). Psychology Press.

Reyna, V. F., Chick, C. F., Corbin, J. C., & Hsia, A. N. (2014). Developmental reversals in risky decision making: Intelligence agents show larger decision biases than college students. *Psychological Science*, *25*(1), 76-84.

Rogers, Y. (2004). New theoretical approaches for HCI. *Annual Review of Information Science and Technology*, *38*(1), 87-143.

Rosenthal, R. E. (1985). Concepts, theory, and techniques principles of multiobjective optimization. *Decision Sciences*, *16*(2), 133-152.

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*(1), 218-231.

Schoemaker, P. J. (1982). The expected utility model: Its variants, purposes, evidence and limitations. *Journal of Economic Literature*, *20*(2), 529-563.

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, *69*(1), 99-118.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*(2), 129.

Simon, H. A. (1979). Rational decision making in business organizations. *The American Economic Review*, *69*(4), 493-513.

Simon, H. A. (1989). The scientist as problem solver. In D. Klahr & K. Kotovsky (Eds.), *Complex information processing: The impact of Herbert A. Simon*, 375-398.

Simon, H.A. (1990). Bounded Rationality. In J. Eatwell, M. Milgate, & P. Newman (Eds.), *Utility and Probability* (pp. 15-18). Palgrave Macmillan London.

Simon, H. A. (1997). *Models of bounded rationality: Empirically grounded economic reason* (Vol. 3). MIT Press.

Schnotz, W., & Kürschner, C. (2007). A reconsideration of cognitive load theory. *Educational Psychology Review*, *19*, 469-508.

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

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, *4*(4), 295-312.

Sweller, J. (2005). Implications of Cognitive Load Theory for Multimedia Learning. In R. E. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning* (pp. 19–30). Cambridge University Press.

Sweller, J., Van Merriënboer, J. J., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, *10*, 251-296.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2*, 53-55

Thiele, L., Miettinen, K., Korhonen, P. J., & Molina, J. (2009). A Preference-Based Evolutionary Algorithm for Multi-Objective Optimization. *Evolutionary Computation, 17*(3), 411-436.

Tractinsky, N. (2018). The usability construct: a dead end?. *Human–Computer Interaction*, *33*(2), 131-177.

Tversky, A. (1975). A critique of expected utility theory: Descriptive and normative considerations. *Erkenntnis*, 163-173.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207-232.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185(*4157*)*, 1124-1131.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293.

Van der Linden, D., Frese, M., & Meijman, T. F. (2003). Mental fatigue and the control of cognitive processes: effects on perseveration and planning. *Acta Psychologica*, *113*(1), 45-65.

Vermeeren, A. P., Law, E. L. C., Roto, V., Obrist, M., Hoonhout, J., & Väänänen-Vainio-Mattila, K. (2010). User experience evaluation methods: current state and development needs. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries* (pp. 521-530). Association for Computing Machinery.

Viale, R. (2020). Why bounded rationality? In R. Viale (Ed.), *Routledge Handbook of Bounded Rationality* (pp. 1-54). Routledge.

Westen, D., & Rosenthal, R. (2003). Quantifying construct validity: two simple measures. *Journal of Personality and Social Psychology*, *84*(3), 608.

White, D. J. (1990). A bibliography on the applications of mathematical programming multiple-objective methods. *Journal of the Operational Research Society*, *41*(8), 669-691.

Wierzbicki, A. P. (1980). The use of reference objectives in multiobjective optimization. In *Multiple Criteria Decision Making Theory and Application: Proceedings of the Third Conference Hagen/Königswinter* (pp. 468-486). Springer Berlin Heidelberg.

Wierzbicki, A. P. (1982). A mathematical basis for satisficing decision making. *Mathematical modelling*, *3*(5), 391-405.

Wierzbicki, A. P. (1999). Reference point approaches. In T. Gal, T.J. Stewart & T. Hanne (Eds.), *Multicriteria Decision Making: Advances in MCDM Models, Algorithms, Theory, and Applications* (pp. 237-275). Springer New York.

Wobbrock, J. O., & Kientz, J. A. (2016). Research contributions in human-computer interaction. *Interactions*, *23*(3), 38-44.

Xin, B., Chen, L., Chen, J., Ishibuchi, H., Hirota, K., & Liu, B. (2018). Interactive multiobjective optimization: A review of the state-of-the-art. *IEEE Access*, *6*, 41256-41279.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, *99*(3), 432-442.