Received: 3 April 2021 | Revised: 3 September 2021 | Accepted: 2 November 2021

DOI: 10.1002/mcda.1773

RESEARCH ARTICLE

WILEY

How do people aggregate value? An experiment with relative importance of criteria and relative goodness of alternatives as inputs

Ulla Ahonen-Jonnarth[1] | Hanna Andersson[1,2] | Fredrik Bökman[1]

[1]Department of Computer and Geospatial Sciences, University of Gävle, Gävle, Sweden

[2]Department of Building Engineering, Energy Systems and Sustainability Science, Faculty of Engineering and Sustainable Development, University of Gävle, Gävle, Sweden

Correspondence
Ulla Ahonen-Jonnarth, Department of Computer and Geospatial Sciences, University of Gävle, Gävle, Sweden.
Email: ulla.ahonen-jonnarth@hig.se

Abstract

The concept of importance of criteria is used as a central element in several decision making contexts, specifically in value aggregation, e.g. as an input to decision support tools. For example, in the analytic hierarchy process (AHP) decision makers are asked to estimate how much more important one criterion is than another. However, it is not clear how people understand aggregation models based on importance of criteria in decision making situations. The purpose of this descriptive study is to investigate if people find an aggregation model in simple value aggregation tasks which remind of the way AHP elicits the input. Further, the purpose is to investigate if people's tendency to find a model depends on their cognitive abilities. In an exploratory laboratory experiment, participants assessed which of two alternatives is the best, based on information about the importance of two criteria and how good the two alternatives are compared to each other with respect to these criteria. The results confirm that people are willing to use importance of criteria and goodness of alternatives as input in value aggregations and show three main models for aggregation. More participants with higher numeracy applied a clear model compared to those with lower numeracy. None of the identified models was one of AHP's models but one of them reminded of one of the ways input can be aggregated in the AHP. The three models identified in the experiment are based on lexicographic order, multiplication and a combination of multiplication and addition. How the results could be used in a prescriptive context is discussed in the paper.

KEYWORDS
multi-criteria aggregation, numeracy, weights of criteria, working memory capacity

1 | INTRODUCTION

It is common that people make statements about importance of criteria in different contexts. Importance of criteria is also used as a central element in decision making and value aggregation, for example, as an input to many decision support tools. One example is the analytic hierarchy process (AHP), one of the most commonly used models in decision analytical support systems, that uses an additive aggregation to rank alternatives (Saaty, 2010). The inputs for the AHP calculations are pairwise comparisons performed by a

decision maker. These comparisons are answers to the following kind of questions: "Which one of criterion 1 and 2 is the more important? How much more important?". AHP uses a 9-graded scale for the judgments of ratios of importances, where 1 means that the two compared criteria are equally important and 9 means that their importances are extremely different. Similarly, the decision maker compares the alternatives pairwise by answering questions of the following type: "Which one of alternatives A and B is better with regard to criterion 1? How much better?". Importance of criteria is also used for additive aggregation in other models than the AHP. For example, a new Best-worst multi-criteria decision making method is based on comparisons of the best criterion and the worst criterion with all other criteria (Rezaei, 2015) instead of all combinations of comparisons between criteria as in the AHP. Another example is the classic procedure Simple Multi-Attribute Rating Technique SMART (Edwards, 1977), where weight coefficients are assessed by letting a decision maker state how much more important each criterion is than the least important. Although SMART has been further developed to SMARTS and SMARTER which take the attribute levels into consideration and do not have importance of criteria as a basic element (Edwards & Barron, 1994), applications of procedures similar to SMART, where weight coefficients in an additive aggregation are assigned without taking the ranges of attribute levels into account, are still abundant.

Importance of criteria is used in several decision making situations and domains. In the context of public procurement, the relative weighting needs to be given for each contract award criterion and this weighting is connected to the importance of criteria (EU, 2014). If the weights are not assigned to different criteria, the descending order of importance of criteria must be given (EU, 2014). A typical example is that price is assigned a weight 0.8 and quality (that can include several aspects) is assigned a weight 0.2 early in the process of a public procurement case, based on a statement that price is four times as important as quality.

In the medical decision making domain, an example of a decision support tool is Annalisa that includes questions of importance of criteria and applies a simple weighted-sum principle for aggregation (Dowie et al., 2013). It is not unusual that weights and valuations of alternatives are assigned by different stakeholders when using Annalisa. For example, in Salkeld et al. (2016) patients using Annalisa assign the relative importance of weights but the assignment of ratings (corresponding the values in aggregation) is based on medical statistics. In the area of Geographic Information Science, AHP is commonly used only partly, for assigning weight coefficients for a weighted sum aggregation without taking the ranges of aspect values into consideration. An example of this is the study of Höfer et al. (2016) where AHP was applied to aggregate different stakeholders' answers about importance of criteria to the weight coefficients.

Thus, importance of criteria is a widely used concept in decision making. One reason for the popularity of its use may be that it makes it easier to handle complex decision problems by dividing them into smaller parts and later aggregating these parts to a solution, as when assignments of scores for alternatives and weights for criteria are separated. However, answering questions about importance of criteria is not unproblematic (as discussed by, e.g., Belton and Gear (1997) and Hämäläinen and Salo (1997)), and Keeney (1992) calls general statements about importance of the objectives, that is, meaning criteria, *the most common critical mistake* in decision making. One example Keeney (1992) takes up as a critical mistake is the question whether the *cost* or the *pollutant concentration* is more important in the context of air pollution regulation. Most people would willingly answer that question and even answer how much more important one of the criteria is without knowing the levels of air pollutant concentration or the costs involved. Keeney (1992) emphasizes that it is necessary to take the specific context and the actual criteria levels, i.e. the actual values of criteria with respect to each alternative, into consideration: How much is the air pollutant concentration reduced and to what cost? Riabacke et al. (2012) have reviewed weight elicitation methods. According to most of the methods they present, weights of criteria can be assigned without taking either the alternatives, their descriptive aspect levels, or the utility levels (utility differences) into consideration.

The use of statements of weight and importance in general and their relation to weight coefficients in particular is, thus, an important problem area that plays a role in theories of aggregation and for decision support tools. Choo et al. (1999) present different ways to interpret criteria weights and conclude that criteria importance is one of the most common interpretations, together with criteria trade-off and scaling factor. Roy and Mousseau (1996) construct a theoretical framework to analyze how relative importance of criteria is taken into account in different aggregation procedures. In an experiment, Korhonen et al. (2013) first asked participants to decide which one of two criteria, European Credit Trading System (ECTS) credit points or Grade Point Average (GPA) is the more important one for next semester studies. After this the participants were asked to choose the best one in pairwise choice alternatives having different levels of ECTS credit points and GPA. An example task for participants in Korhonen et al. (2013): "The first criterion is in terms of ECTS credits and the second in GPA, which one do you prefer: (40, 75) vs. (50, 60)?" The results were used to estimate the weights for criteria so that the choices could be explained using linear value functions and an additive aggregation model. The authors concluded their results by calling into question the statement that the weights reflect the importance of criteria.

It seems to be easy for people to accept the concept of importance of criteria. Is this because the concept is used in many contexts which may lead people to think that they understand what it means? Do people have an idea of a model for how importance of criteria is used in value aggregation or do they find a model when working with value aggregation problems? Some of the differences in how people find and use a model might be partly explained by their cognitive abilities. One such cognitive ability is working memory capacity (WMC) which is of importance in many cognitive tasks (Sörqvist et al., 2010), including reasoning skills (Fletcher et al., 2011), reading comprehension (Engle, 2002) and problem

solving (Conway et al., 2005). When performing a complex task, it is crucial to maintain relevant information in memory as well as a high level of concentration while inhibiting irrelevant stimuli (Fletcher et al., 2011). Individuals with lower WMC usually make more errors in cognitive tasks and have a harder time to focus on the relevant information. Numeracy is another cognitive ability which may be of relevance. Numeracy is defined as "the ability to understand and use numbers" by Reyna et al. (2009) and has been shown to influence people's performance in different judgment and decision making tasks (Cokely et al., 2012; Lindskog et al., 2015). Numeracy has also been observed to play a role when people make probability judgments (Peters et al., 2006; Winman et al., 2014) and interpret statistical concepts describing performance of prediction models (Weissman et al., 2018). Further, people with a lower numeracy are more sensitive to framing effects (Peters et al., 2006) and rely more on heuristics that favors options with lower risk (Cokely & Kelley, 2009). In a review of skilled human decision making in experts and non-experts, Cokely et al. (2018) conclude that numeracy has a higher impact on decision making skills than other more general cognitive abilities, such as cognitive reflection and intelligence. However, people with higher numerical abilities were more or equally susceptible to various decision paradoxes in a conceptual replication study of the psychological phenomena supporting prospect theory (Millroth et al., 2019).

The purpose of this paper is to study how participants perform tasks of aggregating statements of relative importance of criteria and statements of relative goodness (performance) of alternatives with respect to criteria into a judgment of which alternative is best. Answers to these kinds of questions are used as input in some decision support systems, for example the AHP. Our goal is to investigate if and how people without training in decision analysis perform aggregations and make practical use of statements regarding the concepts of importance of criteria and goodness of alternatives. Further, the purpose is to investigate if participants use a clear model for aggregation, and if the use of a model differs with levels of the participants' cognitive abilities numeracy and working memory capacity. Our approach is descriptive, focusing on some types of input that are used in tools aiming to help people in decision making. How the results could be used in a prescriptive context is discussed in section 5 of the paper.

2 | TASK INPUT AND SOME AGGREGATION MODELS

A simple structure for decision problems was constructed to represent specific decision problems by a few statements. This structure involves the following relations:

$$R_1(A,B) = r, r \geq 1$$

$$R_2(B,A) = s, s \geq 1$$

$$R_w(1,2) = t, t \geq 1$$

where $R_1(A,B)$ is the ratio of goodness between two alternatives expressed as a statement of how much better the first argument, alternative A, is compared to the second argument, alternative B, with respect to criterion 1.

$R_2(B,A)$ has the same form as R_1, but with B better than A with respect to criterion 2.

$R_w(1,2)$ is the ratio of importance between the two criteria expressing a statement of how much more important the first argument, criterion 1, is compared to the second argument, criterion 2.

Let us look at examples of statements based on this structure. If we have $R_1(A,B) = 2$,

$R_2(B,A) = 5$, and $R_w(1,2) = 3$ we can formulate statements for several decision problems. A concrete example could be:

Car A is twice as good as Car B with respect to price.

Car B is five times as good as Car A with respect to comfort.

Price is three times as important as comfort.

Is one of the cars A and B overall best, and if so, which one? Or are they equally good?

It is a question of which of the alternatives is the best one or are they equally good ceteris paribus, that is, all other things being equal between the alternatives.

It is also possible to present the problem in an abstract form, to focus on the structure of the decision problem. In the current study, abstract input statements were used instead of concrete alternatives and criteria. The reason for using abstract task descriptions was to ensure that the participants concentrated on the structures of the decision problems and not on the details of concrete alternatives or criteria. In pilot experiments we noticed that concrete alternatives and criteria sometimes affected the answers of the participants because the participants started to discuss their own opinions about the criteria presented, disregarding the actual statements of importance in the experiment. The use of abstract alternatives was chosen to eliminate this kind of effect. Examples of the abstract task statements used are given in the Experimental section.

Input in the form of statements based on $R_1(A,B)$, $R_2(B,A)$, and $R_w(1,2)$ can be aggregated in many different ways. Before the experiment we constructed different models that could potentially be used by the participants. Some of the constructed models were based on literature, some were based on results from pilot experiments, and some were based on our own estimations of how the participants could perform the aggregation. The pre-constructed models were used in the identification of which models the participants used in the experiment, but the models were not presented to the participants. Here we present five of the possible models namely A1 (AHP, distributive mode), A2 (AHP, ideal mode), SM (simple multiplication), MA (multiplication and addition), and Lex (lexicographic). A few other possible models were constructed but these were not used by any of participants, and they are not presented here.

The models, except the lexicographic model, are based on the idea that two numerical values, measures, V_A and V_B, are calculated

representing 'the goodness', performance or the value of alternatives A and B, respectively.

If $V_A > V_B$ then alternative A is better than alternative B.

If $V_B > V_A$ then alternative B is better than alternative A.

If $V_A = V_B$ then alternative A and alternative B are equally good.

2.1 | Model AHP, distributive mode (A1)

The distributive mode of the analytic hierarchy process (AHP) is the first mode published in the AHP context (Saaty, 1987). It uses proportions as inputs, for example, "How much more important is criterion X than criterion Y?" and "How much more important is alternative A than alternative B with regard to criterion X?". In AHP the decision maker makes pairwise comparisons of alternatives with respect to each criterion and also pairwise comparisons of the criteria with respect to the overall goal of the decision problem. From the set of pairwise comparisons of the criteria a matrix is constructed, from which a vector with priorities for the criteria can be calculated. In general, the pairwise judgments may not be perfectly consistent with each other, and Saaty's solution to this was to use a numerical eigenvalue method, and to only accept judgments leading to an inconsistency below a predetermined level. The pairwise comparisons of all alternatives with respect to criterion 1 are treated the same way, and so on. In a final step, the calculated priority vectors are combined to overall priorities for the alternatives by a weighted summation. However, in this example with only two criteria and two alternatives, no inconsistencies are possible, and it is therefore possible to calculate local priorities for each matrix of pairwise comparisons from any of the column vectors, after proper normalization (see Saaty, 2016, p. 368]).

In model A1, corresponding to AHP:s distributed mode, the normalization is with respect to the average. An example of the distributive mode of AHP with proportions $R_1(A,B) = 2$, $R_2(B,A) = 5$ and $R_w(1,2) = 3$ as input follows.

Comparison of criteria

	Criterion 1	Criterion 2	Normalized priorities
Criterion 1	1	3	$\frac{3}{4} = w_1$
Criterion 2	$\frac{1}{3}$	1	$\frac{1}{4} = w_2$
Sum		4	1

Comparison of alternatives with respect to criterion 1

	Alternative A	Alternative B	Normalized priorities
Alternative A	1	2	$\frac{2}{3} = v_{A1}$
Alternative B	$\frac{1}{2}$	1	$\frac{1}{3} = v_{B1}$
Sum		3	1

Comparison of alternatives with respect to criterion 2

	Alternative A	Alternative B	Normalized priorities
Alternative A	1	$\frac{1}{5}$	$\frac{1/5}{6/5} = \frac{1}{6} = v_{A2}$
Alternative B	5	1	$\frac{1}{6/5} = \frac{5}{6} = v_{B2}$
Sum		$\frac{6}{5}$	1

To calculate the total values of alternatives, V_A and V_B, which are called priorities in the AHP-context, the normalized priorities are used as weight coefficients and values of alternatives with respect to each criterion.

$$V_A = w_1 \cdot v_{A1} + w_2 \cdot v_{A2} = \frac{3}{4} \cdot \frac{2}{3} + \frac{1}{4} \cdot \frac{1}{6} = \frac{13}{24} = 0.542$$

$$V_B = w_1 \cdot v_{B1} + w_2 \cdot v_{B2} = \frac{3}{4} \cdot \frac{1}{3} + \frac{1}{4} \cdot \frac{5}{6} = \frac{11}{24} = 0.458$$

Since $V_A > V_B$, alternative A is the best one according to model A1. It is also possible to express V_A and V_B in the following way:

$$V_A = \frac{R_w(1,2)}{R_w(1,2)+1} * \frac{R_1(A,B)}{R_1(A,B)+1} + \frac{1}{R_w(1,2)+1} * \frac{1}{R_2(B,A)+1}$$

$$V_B = \frac{R_w(1,2)}{R_w(1,2)+1} * \frac{1}{R_1(A,B)+1} + \frac{1}{R_w(1,2)+1} * \frac{R_2(B,A)}{R_2(B,A)+1}$$

2.2 | Model AHP, ideal mode (A2)

Another aggregation model in the AHP context is the ideal mode that was constructed in order to prevent rank reversal (see e.g., Saaty, 1999), that is, the change in preference order that may occur for example after addition of a duplicate alternative (Belton & Gear, 1983). The input required from the decision maker is the same in AHP:s distributive and ideal modes, but the normalizations in the calculations of priorities differ. AHP:s ideal mode (our model A2), is similar to AHP:s distributed mode (our model A1), in all except the normalization step. To overcome problems with rank reversal, the ideal mode uses normalization with respect to the largest element, instead of with respect to the average. An example of the ideal mode of AHP with proportions $R_1(A,B) = 2$, $R_2(B,A) = 5$ and $R_w(1,2) = 3$ as input follows.

Comparison of criteria

	Criterion 1	Criterion 2	Normalized priorities
Criterion 1	1	3 (largest)	$\frac{3}{3} = 1 = w_1$
Criterion 2	$\frac{1}{3}$	1	$\frac{1}{3} = w_2$

Comparison of alternatives with respect to criterion 1

	Alternative A	Alternative B	Normalized priorities
Alternative A	1	2 (largest)	$\frac{2}{2}=1=v_{A1}$
Alternative B	$\frac{1}{2}$	1	$\frac{1}{2}=v_{B1}$

Comparison of alternatives with respect to criterion 2

	Alternative A	Alternative B	Normalized priorities
Alternative A	1	$\frac{1}{5}$	$\frac{1}{5}=v_{A2}$
Alternative B	5	1 (largest)	$1=v_{B2}$

As with model A1, to calculate the total values of alternatives, V_A and V_B, the normalized priorities are used as weight coefficients and values of alternatives with respect to each criterion.

$$V_A = w_1 \cdot v_{A1} + w_2 \cdot v_{A2} = 1 \cdot 1 + \frac{1}{3} \cdot \frac{1}{5} = \frac{16}{15}$$

$$V_B = w_1 \cdot v_{B1} + w_2 \cdot v_{B2} = 1 \cdot \frac{1}{2} + \frac{1}{3} \cdot 1 = \frac{5}{6}$$

Normalized values:

$$V_A = \frac{\frac{16}{15}}{\frac{16}{15}+\frac{5}{6}} = 0.561$$

$$V_B = \frac{\frac{5}{6}}{\frac{16}{15}+\frac{5}{6}} = 0.439$$

Since $V_A > V_B$, alternative A is the best one according to model A2. It is also possible to express V_A and V_B, in the following way:

If

$$R_1(A,B) > 1, R_2(B,A) > 1 \text{ and } R_w(1,2) > 1$$

then

$$V_A = 1 + \frac{1}{R_w(1,2) * R_2(B,A)}$$

$$V_B = \frac{1}{R_1(A,B)} + \frac{1}{R_w(1,2)}$$

2.3 | Model of multiplication and addition

Model MA is based on the idea that the value V_i of an alternative i is the sum of two products of ratios and has similarities with model A2. The difference is that normalization is made with respect to the smallest element.

An example of MA with proportions $R_1(A,B) = 2$, $R_2(B,A) = 5$ and $R_w(1,2) = 3$ as input follows. In the tables below, the smallest elements in column 2 is used for normalization and is assigned a value 1 (in model A2 it is the largest element that is assigned the value 1).

Comparison of criteria

	Criterion 1	Criterion 2	Weight
Criterion 1	1	3	$3=w_1$
Criterion 2	$\frac{1}{3}$	1 (smallest)	$1=w_2$

Comparison of alternatives with respect to criterion 1

	Alternative A	Alternative B	Value
Alternative A	1	2	$2=v_{A1}$
Alternative B	$\frac{1}{2}$	1 (smallest)	$1=v_{B1}$

Comparison of alternatives with respect to criterion 2

	Alternative A	Alternative B	Value
Alternative A	1	$\frac{1}{5}$ (smallest)	$\frac{1/5}{1/5}=1=v_{A2}$
Alternative B	5	1	$\frac{1}{1/5}=5=v_{B2}$

The weights and values for each alternative with respect to each criterion are used to calculate the total values of alternatives.

$$V_A = w_1 \cdot v_{A1} + w_2 \cdot v_{A2} = 3 \cdot 2 + 1 \cdot 1 = 7$$

$$V_B = w_1 \cdot v_{B1} + w_2 \cdot v_{B2} = 3 \cdot 1 + 1 \cdot 5 = 8$$

Normalized values:

$$V_A = \frac{7}{7+8} = 0.467$$

$$V_B = \frac{8}{7+8} = 0.533$$

Since $V_A < V_B$, alternative B is the best one according to model MA.

We observed model MA in pilot studies and in the study reported here. For example, a hand-written note from a participant could reveal that calculations were made as

$$V_A = 3 \cdot 2 + 1 = 7$$

and

$$V_B = 3 + 5 = 8$$

when proportions $R_1(A,B) = 2$, $R_2(B,A) = 5$ and $R_w(1,2) = 3$ were used as input.

2.4 | Model of simple multiplication

We observed model SM in pilot studies and in the study reported here. For example, a hand-written note from a participant could reveal that when

$$R_1(A,B) = 2, R_2(B,A) = 5 \text{ and } R_w(1,2) = 3,$$

calculations were made as

$$V_A = 3 \cdot 2 = 6 \text{ and } V_B = 5.$$

The calculations in model SM can be expressed as

$$V_i = R_w(X,Y) * R_X(N,M) \text{ when } R_w(X,Y) > 1$$

and otherwise

$$V_i = R_X(N,M)$$

Note that the value V_i assigned to alternative i is based on only one aspect. If we assume that the model implicitly sets $R_w(Y,X)$ to 1 when $R_w(X,Y) > 1$ the total values of the alternatives, that is, V_A and V_B, are calculated in the following way:

$$V_A = R_w(1,2) * R_1(A,B)$$

$$V_B = R_w(2,1) * R_2(B,A).$$

When $R_1(A,B) = 2$, $R_2(B,A) = 5$ and $R_w(1,2) = 3$, as above, the calculations give that $V_A = 6$ and $V_B = 5$ (normalized values $V_A = 0.545$ and $V_B = 0.455$) and thus alternative A is the best one according to model SM.

2.5 | Lexicographic model (Lex)

The lexicographic model (Lex) is based on the concept of lexicographic order (see, e.g., Fishburn, 1974, Roy & Mousseau, 1996). According to model Lex the alternative that is the best one with respect to the criterion that is stated to be the most important criterion is also totally the best alternative.

The following comparisons can be used to determine if alternative A or B is the best alternative according to Lex, or if they are equally good.

If

$$R_w(1,2) > 1 \text{ and } R_1(A,B) > 1$$

then alternative A is better than B.

If

$$R_w(1,2) > 1 \text{ and } R_1(B,A) > 1$$

then alternative B is better than A.

When $R_1(A,B) = 2$, $R_2(A,B) = 5$ and $R_w(1,2) = 3$, alternative A is the best one according to model Lex since $R_w(1,2) > 1$ and $R_1(A,B) > 1$.

If $R_w(1,2) = 1$, there are different possibilities how to perform the judgment of which of alternatives A and B is the better one. One way to do is to use information about $R_1(X,Y)$ and $R_2(Y,X)$ if $R_w(1,2) = 1$. If $R_1(X,Y) > R_2(Y,X)$ then the alternative X is the best one. Another way is a single-step, or strict, lexicographic model. According to this sub-model of Lex the alternatives are equally good if $R_w(1,2) = 1$, irrespective how they compare with respect to the criteria.

3 | EXPERIMENT

3.1 | Participants

A total of 30 participants (mean age = 27 years, s.d. = 10; 43% women) were recruited at a Swedish university and participated under informed consent, receiving a small honorarium. Written and oral instructions were given in Swedish, and all participants were fluent in Swedish. The educational backgrounds reported by the participants were diverse. About 50% of the participants were students at the university at the time of participating in the experiment. The experimental session lasted between 30 and 120 min, depending on participant. Two of the participants' answers were excluded because of incomplete data.

3.2 | Procedure and materials

The experiment contained five parts, see Table 1. The first part of the experiment was the first aggregation block containing eight questions. The questions in the block were the same for all participants, but the order was randomized. The second part, the second aggregation block, contained 11 questions and half of the participants received the first seven questions in a sequence (from now on referred to as the *sequence condition*). The four last questions were not a part of the sequence. The other group answered the same questions, but the order of the questions was randomized across the second and third aggregation block, parts 2 and 5 in Table 1, respectively (from now on referred to as the *randomized condition*). The third part was the Berlin numeracy test (BNT) by Cokely et al. (2012) and the 4-item paper-and-pencil version was used, although the questions were presented on a computer screen. We used the Swedish translation of the BNT-test validated by Lindskog et al. (2015). After the BNT-test, the participants performed the fourth part that was the size-comparison span task (SIC-span) to estimate their working memory capacity (WMC) (Sörqvist et al., 2010). The fifth part consisted of the third aggregation block with seven questions of the same type as in the first two aggregation blocks.

The participants answered the questions in the experiment using paper and pen in some blocks and using a computer in other blocks (Table 1). Each question about importance of criteria and goodness of alternatives was presented on a separate sheet of paper and the participants were encouraged to write down notes and comments. Further, after each of the three aggregation blocks, the participants were asked to describe in writing the approach they had used to answer the questions.

3.3 | Tasks

The participants answered the questions of which of two alternatives, A or B, is the best one or if they are equally good, given the information about the alternatives and two criteria, 1 and 2.

For example, if we have

$$R_1(A,B) = 2$$

$$R_2(B,A) = 5$$

$$R_w(1,2) = 3$$

the following information and questions were given to the participants:

Suppose that there are two alternatives (A and B) which differ from each other with respect to two criteria (1 and 2). There are no other differences between the alternatives.

A is twice as good as B with respect to criterion 1.

B is five times better than A with respect to criterion 2.

Criterion 1 is three times more important than criterion 2.

Is one of the alternatives A and B overall best, and if so, which one? Or are they equally good?

However, criterion 1 was not always presented as more important than criterion 2 to obtain variation in the questions. In cases where $R_w(2,1) > 1$ instead of $R_w(1,2) > 1$, the rest of the input was changed in a way that the questions always had the same structure. The corresponding reversed question is:

Suppose that there are two alternatives (A and B) which differ from each other with respect to two criteria (1 and 2). There are no other differences between the alternatives.

A is five times better than B with respect to criterion 2.

B is twice as good as A with respect to criterion 1.

Criterion 2 is three times more important than criterion 1.

Is one of the alternatives A and B overall best, and if so, which one? Or are they equally good?

and thus,

$$R_1(B,A) = 5,\ R_2(A,B) = 2,\ R_w(2,1) = 3.$$

We have assumed that participants interpret 'five times better' and 'five times as good as' (in Swedish) as synonyms. It is possible that, in some languages, 'five times better' could be interpreted as 'six times as good as'. However, this interpretation is not plausible in Swedish and it is unlikely that questions like 'x times better' in this study were interpreted as 'x+1 times as good as'. When we analyzed the data to check this, we did not find any support for the latter interpretation.

Most of the questions were constructed to be used in the identification of which model the participant used. These questions are called main questions (see Table 2). Some of the main questions formed a sequence to test serial consequence. In addition, two kinds of control questions were used: control questions with a correct answer due to dominance and consistency questions that were duplicate questions in the experiment.

3.3.1 | Sequences

If statements having a structure as in section 3.3 are presented to a person, it is reasonable that he or she agrees with the following, irrespective how he or she aggregates the input data:

If the participant finds that A is better than B when

$$R_1(A,B) = r, r > 1,$$

$$R_2(B,A) = s, s > 1,$$

$$R_w(1,2) = t, t \geq 1,$$

then [s]he also finds A to be the best alternative in all cases when $R_1(A,B)$ is increased but $R_2(B,A)$ and $R_w(1,2)$ are not changed, that is, when $R_1(A,B) > r$, $R_2(B,A) = s$, and $R_w(1,2) = t$. This follows from transitivity: If A is better than B and A' is an improvement of A it follows that A' is better than B. The person finds A to be the best alternative also in the cases when $R_2(B,A)$ is decreased below s, but $R_1(A,B)$ and $R_w(1,2)$ are not changed. Further, the person finds A to be the best alternative if $R_w(1,2)$ is increased above t, but $R_1(A,B)$ and $R_2(B,A)$ are unchanged. If a person follows these principles, his or her approach can be called serial consequent.

TABLE 2 Summary of type of aggregation questions

Type of aggregation question	Description
Main questions	For identification of a possible model
Serial questions	For identifications of serial consequence, a part of the main questions
Consistency questions	Duplicate questions for identification of consistency, a part of the main questions
Control questions	The only questions with a correct answer

TABLE 1 A summary of all parts in the experiment

Part	Type of tasks	Questions	Answer mode
Part 1	Aggregation block 1	8 questions	Pen and paper
Part 2	Aggregation block 2	11 questions	Pen and paper
Part 3	Berlin numeracy test	4 questions	Computer
Part 4	Size-comparison span task	10 tasks	Computer
Part 5	Aggregation block 3	7 questions	Pen and paper

Serial consequence thus means that a participant changes from alternative X to Y being the best one at some specific step, or not at all, in a sequence where one of the input ratios is incremented or decremented but the other two are held constant, without changing back later in the same sequence. In the sequence used in the experiment, alternative A is the best one for all participants in the beginning of the sequence. During the sequence, the change from alternative A to B as the best alternative occurs at different steps depending on the aggregation model used.

The sequence of tasks used in the experiment had $R_1(A,B)$ and $R_w(1,2)$ held constant, while $R_2(B,A)$ was changed between tasks:

$$R_1(A,B) = 2,$$

$R_2(B,A)$ was changed stepwise from 3 to 9, and

$$R_w(1,2) = 3.$$

The participants answered these serial questions either in order in one block (*sequence condition*) or mixed along two blocks (*randomized condition*). With model SM, the switch from A to B occurs at step 5 and with model MA it occurs at step 3. With models A1, A2, and Lex the switch does not occur, that is, alternative A is the best alternative through the whole sequence.

3.3.2 | Consistency questions

Some of the questions were repeated in another block of the experiment in order to see if the participants answered in a consistent way. As an example, we expected that if a participant early in the experiment session had answered that A is better than B for $R_1(A,B) = x$, $R_2(B,A) = y$ and $R_w(1,2) = z$ [s]he would give the same answer to the same set of statements later in the session. Further, some of these consistency questions were reversed as explained above.

3.3.3 | Control questions with dominance

In the test script, we have inserted control questions with a correct answer, for example, tasks where alternative B is better than A with respect to criterion 2 and equally good with respect to criterion 1. In such tasks, astute participants should realize that alternative B dominates alternative A. Alternative B should be chosen irrespective of the model used by the participant. Thus, these questions were of the form $R_i(A,B) = 1$ and $R_j(B,A) > 1$. In these cases, it is obvious that alternative B is better than A because it is better with respect to the criterion j and the alternatives are equally good with respect to the criterion i. Alternative B should be chosen one irrespective of the values of $R_j(B,A)$ and $R_w(1,2)$. Participants answering control questions correctly presumable understand the questions of the experiment better, or are more alert, than participants not answering the control questions correctly. An example of a control question:

Suppose that there are two alternatives (A and B).

When it comes to criterion 1, A and B are equally good.

When it comes to criterion 2, B is three times better than A.

There are no other differences between the alternatives.

Further, criterion 1 is 5 times more important than the criterion 2.

Is one of the alternatives A and B overall best, and if so, which one? Or are they equally good?

In the experiment, we had three such control questions with a correct answer.

3.4 | Data analysis

In most cases, three sources could be used for assessing which aggregation model each participant used, if any. First, we compared the participant's answers (A is best, B is best or A and B are equally good) to the answers according to possible models we had constructed before the experiment. Because there were only three possible answers, a specific answer could be possible according to several models when looking at a single task, but taking all tasks together, the different models could be assessed. Second, we analyzed the explaining notes and comments the participant had written during the experiment. For some participants, these written explanations were detailed, clearly showing the use of one of the models. On the other hand, in some cases, there were very few written explanations, and in all those cases, the participant did not follow any clear model. Third, we checked at which step the participant switched from alternative A to B being the best one when it comes to serial questions (both sequence condition and randomized condition).

Two of the authors assessed the models independently. In all cases except one, the same main model was identified. After a discussion and re-analysis of data both evaluators agreed on the model classification for all participants.

Control questions with $R_1(A,B) = 1$ (or $R_2(B,A) = 1$) were analyzed separately. They were excluded from the model interpretation because in these cases it is obvious which the correct answer is, but mechanically following a model could in some cases lead to a wrong answer.

4 | RESULTS

4.1 | Models

Of the 28 participants, 16 were classified as using one of the preconstructed models, simple multiplication (SM), multiplication and addition (MA) or the lexicographic model (Lex) and 12 participants did not seem to use any clear model systematically (U - unclear), although half of them showed some attempts to use SM or Lex (U* - unclear*). Generally, it is difficult to discover a model or models only from the participant's answers (A is better, B is better, A and B are equally good) to the tasks. Because of that, the basis for model

assessment was not only comparisons of the answers to the pre-constructed models but also answer notes and block notes and the step when best alternative was changed in a sequence (see details in Appendix A). In the answer sheet notes we did not find evidence for any participant using a model we had not thought of in advance. Due to the participants' notes on the answer sheets, we have assigned three participants to sub-models of the main models. The distribution of the participants over the different models is presented in Table 3.

Most of the participants answered consistently to the serial questions, irrespective of if they belonged to the sequence condition or randomized condition (Table 4).

As can be seen in Table 4, the answers to the serial questions were not affected by the order of the questions (Fischer's exact test, $p = 0.58$). None of the three participants who did not show serial consequence in their answers used a clear model.

4.2 | Numeracy and working memory capacity

The results of the numeracy (BNT) and working memory capacity (WMC) tests are summarized in Table 5.

The participants ($n = 28$) performed the Berlin numeracy test (BNT) (Cokely et al., 2012) in a non-adaptive format, that is, the participants answered all four BNT questions. The Berlin Numeracy test has been shown to give less skewed results than other numeracy scales, and respondents are fairly evenly distributed across the range of scores for large samples (Cokely et al., 2012). Further, the Berlin numeracy test has been validated for Swedish populations (Lindskog et al., 2015), and we have used the same questions in Swedish as in that validation.

In our small sample of respondents, the distribution is strongly skew, with half the participants failing to give a correct answer to any of the four BNT questions (Table 6).

When the BNT data are analyzed according to the adaptive scheme suggested by (Cokely et al., 2012), the results are qualitatively similar to that of a USA web panel (Amazon M-Turk), with approximately half the respondents ending up in the group of lowest numeracy (Cokely et al., 2012).

The respondents who used a clear model scored higher on the numeracy test (median = 2, mean = 1.5) than those who did not use a clear model (median = 0, mean = 0.5). A Cochran-Armitage test with model use as a nominal factor and the number of correct answers on the BNT test as an ordinal factor showed a statistically significant trend of increasing model use with increasing numeracy ($p = 0.017$). Although the sample is small, the experiment shows that respondents with higher numeracy use a model to a larger extent than respondents with lower numeracy.

The participants ($n = 27$) completed SIC-span test (Sörqvist et al., 2010) of working memory capacity (mean = 20.48, s.d. = 9.10, range = 1–38). These results are similar to what have previously been reported (Sörqvist et al., 2010; Sörqvist & Rönnberg, 2012).

The participants for which we could identify a model had higher working memory capacity (mean = 23.3, s.d. = 6.7, $n = 15$) than those who did not seem to use a model (mean = 17.0, s.d. = 10.7, $n = 12$), but the difference is not statistically significant (point-biserial $r = 0.35$, $t[25] = 1.86$, $p = 0.074$).

TABLE 3 Number of participants identified to use a model or unclear model with some (U*) or no (U) attempts to use a model

Model	SM	MA	Lex	U*	U
n	9	3	4	6	6

TABLE 4 Answers to testing serial consequence in sequence and randomized condition

	Serial consequence	No serial consequence
Sequence condition	14	1
Randomized condition	11	2

TABLE 6 Number of correct answers in BNT

Number of correct answers	0	1	2	3	4
Number of participants	13	3	9	3	0

TABLE 7 Participants' results in control questions

	All correct	One or several wrong answers
Clear model	12	4
Not clear model	4	8

TABLE 5 Mean, s.d., minimum and maximum values of BNT* and WMC** tests

	BNT				WMC			
	Mean	s.d.	Min	Max	Mean	s.d.	Min	Max
Participants using a clear model	1.5	1.2	0	3	23.3	6.7	10	36
Participants not using a clear model	0.5	0.8	0	2	17.0	10.7	1	38
Total	1.1	1.1	0	3	20.5	9.1	1	38

Note: * $n = 28$, ** $n = 27$.

4.3 | Control questions

In the experiment, we had three control questions with a correct answer (one participant answered two of them; the other participants answered all of them). 16 participants answered all control questions correctly, and a larger proportion of those participants who used a clear model answered all control questions correctly compared to participants who did not use a clear model (see Table 7).

According to Fischer's exact test ($p = 0.034$, one-sided) there is a statistically significant difference between the groups. The respondents who answered the control questions correctly also to a higher degree used identifiable models.

The median value of BNT for those who did not make any errors in control questions was 2 ($n = 16$) and for those who did at least one error was 0 ($n = 12$). Participants with higher numeracy gave correct answers to the control questions significantly more often than those who had lower numeracy (Cochran-Armitage's test, $p = 0.0064$).

4.4 | Gender

No differences were found between male and female participants in the extent to which they used models (Fischer's exact test, $p = 1$), nor with respect to working memory capacity (Welch's test, $p = 0.61$) or numeracy (Cochran-Armitage's test, $p = 0.32$).

5 | DISCUSSION

In this experiment, the focus is on possible models for value aggregation of the input statements, that is, importance of criteria and goodness of alternatives, without an influence of participants own preferences. Thus, the participants were not asked to make their own judgments about specific criteria and alternatives but the input statements for the aggregation tasks concerned abstract decision problems. The reason for using abstract decision problem was to investigate if and, in that case, how people aggregate the input statements that have the same structure as inputs to some decision support tools. Thus, the study differs from other studies focusing on different perspectives of importance of criteria, elicitation of weight coefficients and related questions, such as the range sensitivity (Beattie & Baron, 1991; Stewart & Ely, 1984), global and local interpretation of weight (Goldstein, 1990; Van Ittersum & Pennings, 2012), elicitation of weight coefficients (Pöyhönen & Hämäläinen, 2001) or connecting the judgments of importance of criteria to weights and impact (with different definitions of impact) as Pajala et al. (2019). Our approach is descriptive, and we wanted to find out how people understand the importance of criteria and goodness of alternatives. This kind of questions are asked in AHP and as Belton and Stewart (2002) point out, it is not clear what importance of criteria in the AHP context mean. Both AHP and multiple attribute value/utility theory (MAVT/MAUT) use additive aggregation but while AHP is based on ratio scales and uses importance of criteria for calculation of priorities, MAVT and MAUT apply scale factors that are based on range differences. In addition to AHP and MAVT/MAUT, outranking methods use another meaning for importance of criteria, which can be seen to correspond to a 'voting strength' (Belton & Stewart, 2002).

A limitation with the current study is that the aggregation task presented to the participants in the experiment is small, with only two criteria and two alternatives. This choice was made in order to make each task reasonably easy to overview, comprising only three statements with relative importance or goodness as inputs to the value aggregation. If the task size is increased to, for example, three criteria and three alternatives, there would be 12 pairwise comparisons as input. In general, with m criteria and n alternatives, the number of pairwise comparisons is m(m-1)/2 + mn(n-1)/2. We believe that using these small aggregation tasks is a relevant way to study how people make use of statements of relative importance of criteria and relative goodness of alternatives, that is, if they use systematic models or not, and if people's tendency to find a model depends on their cognitive abilities. In further studies, more complex aggregation tasks could be used. Another limitation is the relatively low number of participants in the study. If the number of participants was higher, it is possible that other models could have been found, beyond the three main models observed in this study.

All participants used the input statements provided to make a judgment of which of the alternatives is the best one, regardless if they clearly used a model or not. This willingness of the participants to perform the aggregations is in line with earlier observations about people accepting the concept of importance of criteria without a connection to criteria levels (e.g., Keeney, 1992). In our study, this applied regardless of participants' numeracy skills. Even those few participants who presented two different models (and used mainly one of them in aggregation tasks) did not comment that there may be problems because the input can be aggregated in several ways. This may partly depend on the experimental situation and the participants' expectations that there is a correct way to answer the questions. However, the participants were encouraged to make notes both on each answer sheet and after each of three aggregation blocks. One reason for this was to give them an opportunity to be critical. Interestingly, one of the participants wrote an interesting comment "All answers can be correct. It is a question of definition." This person also wrote about feelings in the answer notes and did not use any clear model to aggregate the input.

More than half of the participants applied a clear model to aggregate the input statements. Those participants scored higher in numeracy than those who did not apply a clear model. Numeracy seems thus to be a strong indicator towards the ability to find and apply a clear model for the aggregation task, indicated by the statistically significant effects in this study, even though the number of participants was low. Perhaps those with higher numeracy are more capable of constructing and using quantitative models. Previous studies have shown that people with high numeracy are more coherent in their probability judgments in comparison to those with lower numeracy (Winman et al., 2014). People with lower working memory capacity have been observed to have more difficulties to focus on relevant

information and to make more errors in tasks concerning syllogistic reasoning, categorical thinking, and gambling (Fletcher et al., 2011). The participants who used a clear model in our study had a higher working memory capacity (mean = 23.3, s.d. = 6.7) than the group who did not use a clear model (mean = 17.0, s.d. = 10.7), however, this difference was not significant. Whether this trend is statistically significant could be investigated in a future study with a higher number of participants or another experimental design.

The identified models show three different ways how people without training in decision analysis interpret the tasks and make aggregations. In two of the models, SM (simple multiplication) and MA (multiplication and addition), the participants use all input statements to make a comparison between the alternatives. The third model, Lex (Lexicographic model), focuses on the criterion that is stated to be the most important one (Lexicographic aggregation in Roy and Mousseau (1996)) and this model does not include any calculations. The fact that participants used different models is reasonable because there are no generally accepted theoretical foundations for how the aggregation should be done in the tasks used in the experiment. Still, similar questions are asked as a part of an aggregation process for example in the context of public procurement (e.g., EU, 2014) and as input in decision support tools (for example Saaty, 2010). As Keeney (1992) emphasizes, it is necessary to look at criteria levels in order to discuss importance questions in a specific decision problem context. Part of the problem is that it is possible to unambiguously clarify what *three times longer* means but it is not clear what *three times as good as* actually means, and for example, Belton & Stewart, 2002, (p. 114–115) question the general use of ratio statements to elicit or even approximate weight parameters.

Even if a decision support system was not used in this experiment, the results illustrate the importance of very clearly explaining the required input to the users of decision support systems. In that way these descriptive results have relevance for prescriptive purposes. A problem of using importance of criteria and goodness of alternatives is that the statements can be aggregated in many different ways. For example, in the context of AHP, two different aggregation modes have been used, the original one, that is, the distributive mode, and the one to prevent rank reversal phenomenon, that is, the ideal mode (Saaty, 2010). In addition, different scales have been suggested for the AHP method, for example geometric, logarithmic and balanced (for a review, see Ishizaka & Labib, 2011). To use additive aggregation that is based on value differences and applies weight coefficients as scale factors, as in the context of MAVT/MAUT, is not unproblematic either. Even if it is possible to explain what range differences are, it is not necessarily easy to work with them correctly. For example, if ranges are changed people may not change the weight coefficients enough which is known as range insensitivity (Montibeller & von Winterfeldt, 2015). How large the range insensitivity is, depends on the method used for weight elicitation (Fischer, 1995; von Nitzsch & Weber, 1993). Regardless of which aggregation model that is used, the basis of it should be explained to the users so that it is clear for them what kind of input is required and how the input is used in the model of the tool.

According to what we could infer, none of the participants used the input in one of the ways aggregation is performed in the AHP (our models A1 and A2) which indicates that the aggregation models used in the AHP may not be intuitive. If they would be intuitive then we expect that at least a few of the participants would have used them in their aggregation. However, model MA (multiplication and addition, used by 4 participants) could be used as another way to make an AHP-kind of aggregation, having similarities with model A2. In AHP model A2, normalization is made with respect to the largest element in each priority vector (see Section 2.2). In model MA normalization is instead made with respect to the smallest element (see Section 2.3). In the notes made by the participants we saw no signs of calculations using a matrix as in Section 2.3 but they used calculations applying the input values directly, in a way that is consistent with $V_X = w_1 \cdot v_{X1} + w_2 \cdot v_{X2}$.

For example, when $R_w(1,2) > 1$, $R_1(A,B) > 1$ and $R_2(B,A) > 1$
$w_1 = R_w(1,2)$ and $w_2 = 1$ (the smallest element),
$v_{A1} = R_1(A,B)$ and $v_{B1} = 1$ (the smallest element), and
$v_{A2} = 1$ (the smallest element) and $v_{B2} = R_2(B,A)$
leads to

$$V_A = R_w(1,2) * R_1(A,B) + 1$$

$$V_B = R_w(1,2) + R_2(B,A).$$

See an example in Section 2.3.

We could see that four of participants in the study performed calculations of V_A and V_B in this way, in accordance with model MA. However, from this we do not infer that they thought a kind of normalization was involved in their calculations.

As stated by Choo et al. (1999), ratio scales are assumed in AHP whereas multiattribute value function models use interval scales. The statements of importance of criteria and the goodness of the alternatives, used in the current experiment, require the use of ratio scales. As exemplified above, model MA uses the importance of criteria in a way that reminds of AHP:s ideal mode (our model A2). Models SM and MA use importance of criteria as the basis for weight coefficients, even if it is unclear if participants in the experiment thought about it explicitly. However, the use of importance of criteria differs between models SM and MA. Model MA applies an additive aggregation, belonging to the basic group weighted sum in the classification by Roy and Mousseau (1996). Model SM, the most commonly used model in our experiment (used by 9 participants, Table 3), incorporates another interpretation of importance of criteria. The input statements include information of criteria and alternatives on ratio scale, that is, $R_i(X,Y) = r$. Then, $R_i(Y,X) = \frac{1}{r}$. For example, if a distance X is three times as long as another distance Y, then distance Y is $\frac{1}{3}$ as long as X. Similarly, if alternative A is three times as good as B with respect to criterion 1, it would follow that alternative B is $\frac{1}{3}$ times as good as A respect to criterion 1. According to model SM, if $R_w(1,2) = r$, $r \geq 1$, it follows that $R_w(2,1) = 1$, i.e. R_w is not used as a ratio because $R_w(2,1) \neq \frac{1}{r}$. Further, according to model SM the value of an alternative is based on one of the criteria, either with or without a

multiplication with the importance of that criterion (see details in Section 2.4). A value aggregation based on both criteria is thus not a part of this model. It is possible that some participants with high numeracy find a model that seems reasonable and use it without realizing that there are ways to aggregate the input statements using both criteria for each alternative.

For each of models SM, MA, and Lex we observed modified versions. The sub-model SM-m (modified) includes use of a threshold value. One participant explicitly explained that when the difference between the goodness of A and B is low, the alternatives are judged to be equally good. The sub-model Lex-m (modified) is a single-step lexicographic model. This means that only the criterion that is stated to be the most important matters. If the alternatives are equally good with respect to that criterion, they are also equally good in total, irrespective how they compare with respect to other criteria. It is possible to apply the sub-models SM-m and Lex-m unambiguously, if for sub-model SM, a threshold value is defined. On the other hand, a sub-model MA-g (gut feeling) is based on personal valuations and cannot be expressed explicitly, or at least it is difficult to do it. The participant who used model MA-g sometimes calculated the result using model MA but answered something else than the calculated result suggested. In some questions the participant commented that the answer was based on a gut feeling about spreading a risk or thoughts about what might be morally right to do.

The observed high level of serial consequence irrespective if participants belonged to the sequence condition or the randomized condition indicates that the participants had an idea of the kind of rationality serial consequence means, that is, that a change of alternative A or B being the best alternative occurs in one step in a series where two of R_1, R_2, and R_w are kept constant and one is changed. High level of serial consequence also indicates that the participants concentrated on the tasks properly. Only three of the participants did not show serial consequence, and none of them used a clear model in the aggregation tasks. In the series used in our study, $R_1(A,B)$ and $R_w(1,2)$ were kept constant and $R_2(B,A)$ was increased. It is plausible that most of the participants in the sequence condition, who answered the questions in one block in order with increase of $R_2(B,A)$ from 3 to 9, realized that it is reasonable that a change of alternative A to B being the best alternative occurs in one step, or not at all when model Lex was applied (see Appendix A). For participants in the randomized condition, the situation was different, and it is unlikely that the participants that received the sequence tasks randomized over several blocks remembered exactly what questions they had received previously, or their own answers. This gives support to an interpretation that serial consequence follows the models that were used by the participants.

Two kinds of control questions were used: duplicate questions and questions with correct answers. That participants gave the same answers on the duplicate questions in different blocks strengthened the identification of a certain model. In addition, the same answers on these questions indicate that participants used similar way to aggregate the input statements between blocks, that they concentrated during the experiment and that the answers were not produced randomly. In a few cases, the participants did not answer in the same way on one of the duplicate questions even if they used a clear model. In these cases, there were notes that supported the identification of a model strongly, and serial consequence applied. To correctly answer the control questions with correct answers a participant needs to realize that this particular question contains information that most of the other questions do not. The results from the current study suggest that higher numeracy might play a role in realizing what the questions with correct answers mean. This is supported by the finding that participants with higher numeracy more often answered correctly to the control questions than those who had lower numeracy. Using other type of questions, Weissman et al. (2018) found a correlation between correct answers and numeracy in a study of participants' interpretation of statistical concepts describing performance of prediction models.

Even if there are problems with the concept of importance of criteria in a single aggregation, statements of weights could be used in communication in a decision making process, for example when two or several aggregations are compared. If this communication is going to work well, the concept of weight should be understood in a similar way by the persons involved in the process. Odelstad (1990) shows that it is unrealistic to use importance of criteria in a single aggregation, but that it is meaningful to talk of importance of criteria when two different aggregations are compared.

6 | CONCLUSIONS

This explorative study investigated how people aggregate input statements of relative importance of criteria and input statements of relative goodness (performance) of alternatives with respect to criteria into a judgment of which alternative is best. The statements remind of statements that are used in different contexts, including decision support tools such as AHP, the analytic hierarchy process. Further, the study investigated if participants used a clear model for aggregation, and if the use of a model differed with levels of the participants' cognitive abilities numeracy and working memory capacity. Thus the study, an exploratory laboratory experiment, has a descriptive approach. The results show that people are willing to solve simple aggregation tasks with only statements about importance of criteria and goodness of alternatives as input, which is in line with for example Keeney (1992) and Belton and Stewart (2002). Three clearly different models, and a few sub-models, were applied by the participants. In the most commonly used model, SM (simple multiplication), the inputs of importance of criteria were not used as ratios. None of the models identified were one of the AHP modes, but one of the models, MA (multiplication and addition), had similarities with one of the AHP modes. Participants who scored higher in numeracy to a significantly greater extent applied a clear model compared to those who scored lower in numeracy. The facts that it is possible to aggregate the input statements in many ways and that it is not clear how users understand the underlying model makes it problematic to use these kinds of statements as input in a decision support system without clear guidelines of how the input should be given and how it is used in the model of the tool. The results of this

study show, even if the number of participants was relatively low, that people have different ideas about how to aggregate statements about goodness of alternatives and importance of criteria. We argue that it is important for the analyst working prescriptively, to take this into account when working with decision makers or constructing decision support tools. In further studies, more complex aggregation tasks could be used with a larger number of participants.

ACKNOWLEDGMENTS

The authors want to thank Jan Odelstad for valuable discussions and comments. Further, the authors wish to thank two anonymous reviewers for their helpful comments.

DATA AVAILABILITY STATEMENT

Data used for statistical tests is included in Appendix A. Data for model classification, the written notes in answer sheets and block sheets, are not possible to submit, because they are hand written notes (in Swedish) and could be used to identify the participants.

ORCID

Ulla Ahonen-Jonnarth https://orcid.org/0000-0001-9933-8308
Hanna Andersson https://orcid.org/0000-0001-6151-9664
Fredrik Bökman https://orcid.org/0000-0001-5220-9293

REFERENCES

Beattie, J., & Baron, J. (1991). Investigating the effect of stimulus range on attribute weight. *Journal of Experimental Psychology*, *17*, 571–585.

Belton, V., & Gear, T. (1983). On a shortcoming of Saaty's analytic hierarchy process. *Omega*, *11*, 228–230.

Belton, V., & Gear, T. (1997). On the meaning if relative importance. *Journal of Multi-Criteria Decision Analysis*, *6*, 335–338.

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

Choo, E. U., Schoner, B., & Wedley, W. C. (1999). Interpretation of criteria weights in multicriteria decision making. *Computers & Industrial Engineering*, *37*, 527–541.

Cokely, E. T., & Kelley, C. M. (2009). Cognitive abilities and superior decision making under risk: A protocol analysis and process model evaluation. *Judgment and Decision making*, *4*, 20–33.

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin numeracy test. *Judgment and Decision making*, *7*, 25–47.

Cokely, E. T., Feltz, A., Ghazal, S., Allan, J. N., Petrova, D., & Garcia-Retamero, R. (2018). In K. A. Ericsson, R. R. Hoffman, A. Kozbelt, & A. M. Williams (Eds.), *Cambridge handbooks in psychology. The Cambridge handbook of expertise and expert performance* (pp. 476–505). Cambridge University.

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*, 769–786.

Dowie, J., Kaltoft, K. M., Salkeld, G., & Cunich, M. (2013). Towards generic online multicriteria decision support in patient-centred health care. *Health Expectations*, *18*, 689–702.

Edwards, W. (1977). How to use multiattribute utility measurement for social Decisionmaking. *IEEE Transactions on Systems, Man, and Cybernetics*, *7*, 326–340.

Edwards, W., & Barron, F. H. (1994). SMARTS and SMARTER: Improved simple methods for multiattribute Utlity measurement. *Organizational, Behavior and Human Decision Processes*, *60*, 306–325.

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science*, *11*, 19–23.

EU, (2014). Directive 2014/24/EU of the European parliament and of the council of 26 February on public procurement and repealing Directive 2004/18/EC. Official Journal of the European Union, L 94/65

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

Fishburn, P. C. (1974). Exceptional paper—Lexicographic orders, utilities and decision rules: A survey. *Management Science*, *20*, 1442–1471.

Fletcher, J. M., Marks, A. D., & Hine, D. W. (2011). Working memory capacity and cognitive styles in decision-making. *Personality and Individual Differences*, *50*, 1136–1141.

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

Hämäläinen, R. P., & Salo, A. A. (1997). The issue is understanding the weights. *Journal of Multi-Criteria Decision Analysis*, *6*, 340–343.

Höfer, T., Sunak, Y., Siddique, H., & Madlener, R. (2016). Wind farm siting using a spatial analytic hierarchy process approach - a case study of the Städteregion Aachen. *Applied Energy*, *163*, 222–243.

Ishizaka, A., & Labib, A. (2011). Review of the main developments in the analytic hierarchy process. *Expert Systems with Applications*, *38*, 14336–14345.

Keeney, R. L. (1992). *Value-focused thinking: A path to creative decisionmaking*. Harvard University Press.

Korhonen, P. J., Silvennoinen, K., Wallenius, J., & Öörni, A. (2013). A careful look at the importance of criteria and weights. *Annals of Operations Research*, *211*, 565–578.

Lindskog, M., Kerim, N., Winman, A., & Juslin, P. (2015). A Swedish validation of the Berlin numeracy test. *Scandinavian Journal of Psychology*, *56*, 132–139.

Millroth, P., Nilsson, H., & Juslin, P. (2019). The decision paradoxes motivating Prospect theory: The prevalence of the paradoxes increases with numerical ability. *Judgment and Decision making*, *14*, 513.

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

Odelstad, J. (1990). Mätning och beslut. Measurement and Decision. University of Uppsala, Sweden.

Pajala, T., Korhonen, P., & Wallenius, J. (2019). Judgments of importance revisited: What do they mean? *Journal of the Operational Research Society*, *70*, 1140–1148.

Peters, E., Västfjäll, D., Slovic, P., Mertz, C. K., Mazzocco, K., & Dickert, S. (2006). Numeracy and decision making. *Psychological Science*, *17*, 407–413.

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

Reyna, V. F., Nelson, W. L., Han, P. K., & Dieckmann, N. F. (2009). How numeracy influences risk comprehension and medical decision making. *Psychological Bulletin*, *135*, 943–973.

Rezaei, J. (2015). Best-worst multi-criteria decision-making method. *Omega*, *53*, 49–57.

Riabacke, M., Danielson, M., & Ekenberg, L. (2012). State-of-the-art prescriptive criteria weight elicitation. *Advances in Decision Making*, *2012*, 276584.

Roy, B., & Mousseau, V. (1996). A theoretical framework for analysing the notion of relative importance of criteria. *Journal of Multi-Criteria Decision Analysis*, *5*, 145–159.

Saaty, R. W. (1987). The analytic hierarchy process—What it is and how it is used. *Mathematical Modelling*, *9*, 161–176.

Saaty, T. L. (1999). Basic theory of the analytic hierarchy process: How to make a decision. *Revista de la Real Academia de Ciencias Exactas Físicas y Naturales*, *93*, 395–423.

Saaty, T. L. (2010). *Mathematical principles of decision making: Generalization of the analytic network process to neural firing and synthesis*. RWS Publications.

Saaty, T. L. (2016). The analytic Hierachy and analytic network processes. In S. Greco, M. Ehrgott, & J. R. Figueira (Eds.), *Trends in multiple criteria decision analysis* (pp. 363–420). Springer.

Salkeld, G., Cunich, M., Dowie, J., Howard, K., Patel, M. I., Mann, G., & Lipworth, W. (2016). The role of personalised choice in decision support: A randomized controlled trial of an online decision aid for prostate cancer screening. *PLoS One*, *11*, e0152999.

Stewart, T., & Ely, D.W. (1984). Range sensitivity: A necessary condition and a test for the validity of weights. National Center for Atmospheric Research report NCAR 3141–84/14, Boulder, Colorado

Sörqvist, P., Ljungberg, J., & Ljung, R. (2010). A sub-process view of working memory capacity: Evidence from effects of speech on prose memory. *Memory*, *18*, 310–326. https://doi.org/10.1080/09658211003601530

Sörqvist, P., & Rönnberg, J. (2012). Episodic long-term memory of spoken discourse masked by speech: What is the role for working memory capacity? *Journal of Speech, Language and Hearing Research*, *55*, 210–218. https://doi.org/10.1044/1092-4388(2011/10-0353)

Van Ittersum, K., & Pennings, J. M. E. (2012). Attribute-value functions as global interpretations of attribute importance. *Organizational Behavior and Human Decision Processes*, *119*, 89–102. https://doi.org/10.1016/j.obhdp.2012.04.002

Weissman, G. E., Yadav, K. N., Madden, V., Courtright, K. R., Hart, J. L., Asch, D. A., Schapira, M. M., & Halpern, S. D. (2018). Numeracy and understanding of quantitative aspects of predictive models: A pilot study. *Applied Clinical Informatics*, *9*, 683–692. https://doi.org/10.1055/s-0038-1669457

Winman, A., Juslin, P., Lindskog, M., Nilsson, H., & Kerimi, N. (2014). The role of ANS acuity and numeracy for the calibration and the coherence of subjective probability judgments. *Frontiers in Psychology*, *5*, 1–15. https://doi.org/10.3389/fpsyg.2014.00851

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

How to cite this article: Ahonen-Jonnarth, U., Andersson, H., & Bökman, F. (2022). How do people aggregate value? An experiment with relative importance of criteria and relative goodness of alternatives as inputs. *Journal of Multi-Criteria Decision Analysis*, *29*(3-4), 259–273. https://doi.org/10.1002/mcda.1773

APPENDIX A

Basis for model assignments

Basis for model assignments and some specific notes in the table below. Columns:

(2) Model identified. The main models are SM (simple multiplication), Lex (Lexicographic model), MA (multiplication and addition) and U (Unclear, i.e., with no clear model). The sub-models are SM-m (SM-modified), MA-g (MA-gut feeling), and Lex-m (Lex-modified). Half of those who did not use a clear model showed attempts to partly use SM or MA and are classified as U-SM or U-MA, respectively.

(3) For participants with number 1–22 column (3) shows the number of answers according to the model in column 2, that is, SM, MA, or Lex. For participants 23–28 column (3) shows the highest number of answers according to any of the models in the analysis, although this model was not assigned as a clear model to this participant in the assessment. The total number of answers was 23 in most cases. If the total number of answers differs from 23 it is shown in a denominator.

(4) Judgments of two of the authors about how strong support, if any, the answers to the aggregation main questions give to the identified model (SM, MA, Lex, and subgroups): s – strong, m – medium, w – weak.

(5) Column 5 shows when a participant changed from assessing A as being the best alternative to B. For example, according to model SM, this change occurs when $R_2(B,A) = 5$. Those using Lex are marked by A because they never changed the alternative to be the best one from A to B, in accordance with model Lex. Three participants changed their answers back and forth between A and B, which is marked with 'not conseq' in the Table.

(6) Our judgments of how strong support, if any, the block notes give to the classification of the use of a model (SM, MA, Lex, U, and subgroups), with s, m, and w as in column (4).

(7) Our judgments of how strong support, if any, the answer notes give to the classification of the use of a model (SM, MA, Lex, U, and subgroups), with s, m, and w as in column (4).

(8) Score in Berlin numeracy test.

(9) Score in the size-comparison span task (SIC-span) to estimate working memory capacity (WMC).

1. Nr	2. Model	3. Number of answers	4. Model support	5. Serie change	6. Block notes	7. Answer notes	8. BNT score	9. WMC score
1	SM	23	s	5	s		3	10
2	SM	23	s	5	s		3	26
3	SM	23	s	5	s	s	2	29
4	SM	22	s	5	q		0	18
5	SM	23	s	5	w		2	23
6	SM	23	s	5	m		3	29
7	SM	22	s	5	w	m	0	22
8	SM	23	s	5	s	s	2	27
9	SM-m	20/22	s	5	s		2	23
10	MA	23	s	3	s	m	2	36
11	MA	22	s	3	w	w	2	29
12	MA-g	19/22	m	3	m	s	1	14
13	Lex	23	s	A	s	s	2	23
14	Lex	23	s	A	s	s	0	24
15	Lex	20	m	A	s		0	
16	Lex-m	21/22	s	A	s	s	0	16
17	U-SM	19	m	5		s	2	38
18	U-SM	21	m	4	s	m	1	10
19	U-SM	20	m	5	w	w	0	1
20	U-SM	18	m	6	m	m	1	10
21	U-Lex	14		3	s		0	24
22	U-Lex	15	w	5	s	s	0	5
23	U	14		4	s	m	0	14
24	U	16		Not conseq,	s	w	2	15
25	U	18		3		m	0	18
26	U	17		Not conseq,	s		0	13
27	U	8/20		Not conseq,	s		0	29
28	U	21		A	s	s	0	27