Inter- and Intra-Personal Differences, and Consistency of Decision Rules, in Multi-Criteria Modelling Method DEX: A Preliminary Study

Marko Bohanec
Jožef Stefan Institute, Department of Knowledge Technologies
Jamova cesta 39, SI-1000 Ljubljana, Slovenia
and
University of Nova Gorica, Glavni trg 8, SI-5271 Vipava, Slovenia
marko.bohanec@ijs.si

***Abstract**. DEX (Decision EXpert) is a qualitative multi-criteria decision-modelling method in which decision alternatives are evaluated according to decision rules, elicited from individual decision makers. In this preliminary study, we assessed the differences between decision rules acquired from different subjects (inter-personal differences) and from the same subjects at different times (intra-personal differences). We also assessed the consistency of so-acquired rules and the ability of subjects to estimate the importance (weights) of criteria. The results indicate a high variability of decision rules, both inter- and intra-personal. Intra-personal drift is lower than inter-personal differences, but not substantially. The consistency of rules varied between a small decision table with clearly ordered criteria, where it was almost perfect, and a large decision table with less apparent preferential relations, where it was rather poor at the average level of 0.77. Criteria weighs also drifted at the rate 9–19% per month.*

Keywords. Multi-criteria decision modelling, method DEX, decision rules, weights, consistency, drift

1 Introduction

Multi-criteria decision modelling (MCDM) is a decision-making technique that involves the use of models to evaluate a set of decision alternatives based on multiple criteria or objectives (Greco, et al., 2016; Thakkar, 2021). MCDM is used in situations where the decision maker needs to balance the trade-offs between multiple, possibly conflicting criteria. MCDM typically involves three steps: (1) defining the decision problem and criteria, (2) identifying and evaluating the alternatives, and (3) synthesizing the results to make a decision. There are many MCDM methods that differ in how they represent criteria, evaluation/aggregation rules and alternatives, and how they acquire this information, which is often subjective, from decision makers. MCDM methods are typically named using acronyms, such as WSM, AHP, ANP, MAUT/MAVT, TOPSIS, VIKOR, MACBETH, PAPRIKA, PROMETHEE, ELECTRE, UTA, DRSA, DEX; see Greco, et al., (2016), Thakkar (2021) and Kulkarni (2022) for overviews and more information.

In this study, we are particularly interested in aggregation/evaluation aspects of multi-criteria models. In order to evaluate alternatives, a vast majority of MCDM methods employ the weighted sum

$$f(x_1, x_2, \dots, x_n) = w_1 x_1 + \cdots + x_n x_n$$

Here, x_i and w_i denote numerical criteria and their weights, respectively. The larger the weight, the more influential the criterion. Weights $w_1, w_2, \dots, w_n$ are often normalized so that their sum or maximum equals to some predefined number, typically 1 or 100. Generally, weights are subjective and need to be acquired from individual decision makers (Rezaei, et al., 2021; Silva, et al., 2021).

On one hand, MCDM methods strive to obtain weights that as accurately as possible represent decision maker's preferences. A good example is the method AHP (Analytic Hierarchy Process) (Saaty & Vargas, 2012), which proceeds by asking the user to assess relative importance of pairs of criteria, using the scale from 1 (equal importance) to 9 (extreme importance of one criterion over other). On this basis, AHP calculates criteria weights and assesses the consistency of user's information.

On the other hand, weights are subjective. Not only that they differ between different decision makers, they can change ("drift") also with the same person due to changes in the decision context, changes of their preferences or just their inability to express their preferences accurately enough. Thus, the question is how well can we assess criteria weights and what inter- and intra-personal differences should we expect.

This study is aimed at answering these questions in relation with the decision modelling method DEX (Decision EXpert) (Bohanec, 2022). DEX is a

qualitative MCDM method. It is somewhat specific in that it uses qualitative criteria and decision rules. Variables that represent criteria in DEX models are not numeric, but discrete and symbolic, using words as their values instead of numbers. For example, the criterion *Price* can be assessed using three categories "high", "medium", "low", and *Technical characteristic*s of some system can be assessed as "bad", "acceptable", "good", or "excellent". Consequently, in order to evaluate decision alternatives, DEX does not employ the weighted sum, but *decision rules* that take the general form:

if $x_1 = v_1$ and $x_2 = v_2$ and … and $x_n = v_n$
then $f(x_1, x_2, \ldots, x_n) = v_y$

Here, x_i are qualitative criteria and v_i are some categories taken from the corresponding value scales. Similarly as with weights, decision rules are acquired from the decision maker and conveniently represented in terms of *decision tables* (see example in Table 1).

In this study we addressed the following research questions:

A. *Inter-personal differences*: How much do decision tables acquired from different subjects differ?
B. *Intra-personal differences*: How much do decision tables acquired from the same subject at different times differ?
C. *Consistency*: Are decision rules, formulated by subjects, consistent and to which extent?
D. *Weight assessment*: Can subjects, who defined decision rules, also assess the weights of criteria and how well?

This study is considered preliminary because it has so far involved a relatively small number of respondents with narrow backgrounds, mostly students and researchers. Question B turned out particularly difficult, as it would require a well-controlled experimental setup with precise time differences between the trials, which we have not attempted so far.

2 Methods

The methodological approach is based on a questionnaire that asks respondents to define decision rules in two predefined decision tables, one assumingly easy and one more difficult. The experiment aims to exclude any tools that might help the respondents to formulate decision rules, thus the questionnaire is answered on paper. The response time is not limited (but is typically well within the 10-minutes range). All participants thus far were from Slovenia, therefore the questionnaire was formulated in the Slovenian language. An English translation is presented hereafter.

The first task (*Car*) is to define decision rules for evaluating a family car considering just two criteria: *Price* and *Technical characteristics*. An empty decision table consisting of 12 possible combinations of the criteria's discrete values is presented to the respondents, asking them to mark the corresponding values of *Car* (Table 1). In connection with this table, respondents are also asked to assess the weights of the two criteria, as shown in Table 2.

The second task (*Store*) is more difficult; it was inspired by the experiment designed by Vetschera, et al. (2014), but uses a reduced number of categories to keep the decision table reasonably small. The task is to assess the suitability/attractiveness of the store for daily purchases, primarily referring to purchases of groceries. Four qualitative criteria are suggested:

- *Store size*: "market" or "supermarket";
- Walking *distance* from home: "less than 10 minutes", "more than 10 minutes";
- *Price category*: "lower", "higher";
- Product *quality*: "lower", "higher".

There are the same four possible outcomes as with *Car*: "unacc", "accept", "good", and "excel".

Notice that the four *Store* criteria are binary (two-valued); this yields 16 possible value combinations, which are presented in the questionnaire in a table similar to Table 1. An analogous question to that from Table 2 is asked for the *Store* task, too.

The second task is considered more difficult than the first one because it involves twice as many criteria. The questionnaire itself does not interpret the criteria and their values any further, so we may expect subjective individual interpretations. In contrast with *Car*, where all the criteria are clearly preferentially ordered and we may expect that decision rules will reflect this order, this is much less so with *Store*. Buying habits largely differ between consumers, and while one may prefer a "lower" price category, some other may equally well prefer the "higher". Thus, we can hardly expect any clear preferential ordering of rules in the *Store* case.

Table 1: A table for acquiring decision rules for the evaluation of cars.

	Price	Tech.char.	Car			
			unacc	accept	good	excel
1	high	poor				
2	high	accept				
3	high	good				
4	high	excel				
5	medium	poor				
6	medium	accept				
7	medium	good				
8	medium	excel				
9	low	poor				
10	low	accept				
11	low	good				
12	low	excel				

Table 2: Question to assess *Car* weights.

Please assess criteria weights so that their sum equals 100:

Criterion	Weight
Price	
Tech.char.	
Sum	100

2.1 Differences between Decision Tables

Research questions A and B require the calculation of differences between two decision tables. A table T_t can be represented as:

$$T_t = \langle y_{t,1}, y_{t,2}, \dots, y_{t,k} \rangle$$

This is a vector of k ordinal numbers $y_{t,r} \in \{1,2,3,4\}, r = 1, \dots, k$, where r is the rule index, and k equals to 12 and 16 for the *Car* and *Store* tables, respectively. Notice that the lowest and highest possible vectors are $\langle \underbrace{111 \dots 1}_{k} \rangle$ and $\langle \underbrace{444 \dots 4}_{k} \rangle$. This gives the following formula for calculating the difference between two decision tables:

$$\Delta T_a, T_b = \frac{1}{k} \sum_{i=1}^{k} \frac{|y_{a,i} - y_{b,i}|}{3}$$

This formula yields the difference of 1 for the above extreme case, and 0 for two equal decision tables.

2.2 Consistency of Decision Rules

Whenever value scales of all involved criteria are preferentially ordered (from 'bad' to 'good' values or vice versa), we can assume that "rational" decision rules will be *consistent*: the better the input criteria (such as *Price* and *Technical characteristics*), the better the outcome (*Car*). Consequently, the decision table is expected to obey the *principle of dominance*, so that the aggregation function is *monotone*: improving or at least staying constant in the direction of each improving criterion.

In respondents' decision tables we observed and counted decision rules that violated the monotonicity constraint. In principle, well-defined *Car* decision tables are expected to be consistent. This is generally not true for *Store*, which was primarily aimed at assessing the average level of (in)consistency in that case.

2.3 Assessment of Weights from Decision Tables

Even though DEX is a qualitative method, for which the concept of criteria weights is somewhat unnatural, it is possible to approximately assess weights from a defined decision table. In this study, we used three methods. The first two, *Gini gain* (GG) and *Information gain* (IG), are routinely used in machine learning for determining the strength of features from data (Raileanu & Stoffel, 2004; Rokach & Maimon, 2015, pp. 62–63). These methods employ *impurity measures Gini* and *Entropy*, respectively, to assess the disorder of data (i.e., a DEX decision table T):

$$Gini(T) = 1 - \sum_{i=1}^{n} p^2(v_i)$$

$$Entropy(T) = -\sum_{i=1}^{n} p(v_i) \log_2 p(v_i)$$

Here, $p(v_i)$ is the probability/proportion of value v_i occurring in the table and n is the number of input criteria ($n = 2$ for *Car* and 4 for *Store*).

Using these measures, a relative weight RW of some criterion c is determined as

$$RW(c,T) = M(T) - \sum_{v \in S_c} \frac{|T_{c=v}|}{|T|} M(T_{c=v})$$

where M is an impurity measure (*Entropy* for IG and *Gini* for GG), v are values taken from S_c, the qualitative value scale of c, and $T_{c=v}$ denotes the part of T where $c = v$. $|T|$ denotes the size of table T in terms of the number of decision rules,.

The third method, *Linear approximation* (LA), is implemented in the software DEXi (Bohanec, 2020). It interprets decision rules as points in a multi-dimensional space and approximates them with a hyperplane using the least squares principle. Criteria weights are approximated from the slopes of the hyperplane: the higher the slope in the direction of a criterion, the higher the corresponding relative weight of the criterion. For more details about LA, see Bohanec & Zupan (2004, sec. 3.4) and Deguine, et al. (2021, sup. sec. 2).

Among the three methods, LA seems better suited for consistent, monotone and potentially linear decision tables, such as *Car*, while GG and IG might better capture the importance of variables in general data tables, such as *Store*.

2.4 Comparison of Weights

In order to compare weights $W_M = w_1, \dots, w_n$, estimated by some method M from decision rules, and weights $\Omega = \omega_1, \dots, \omega_n$ as given by the respondent, we used the formula:

$$\Delta W_M, \Omega = \frac{1}{n} \sum_{i=1}^{n} \frac{|w_i - \omega_i|}{100}$$

Here, n represents the number of criteria. The range of weights w_i and ω_i is [0,100]. Then, the range of $\Delta W_M, \Omega$ is from 0 (no difference) to 1 (extreme difference). The same formula is applicable for comparing weights obtained by two methods.

3 Preliminary Results

To date (May 2023), 34 participants answered the questionnaire, mostly students, colleagues at work and friends. Among these, 17 were approached about one month later and asked to answer exactly the same questionnaire; this served for the assessment of intra-personal differences between their first and second trials. The total number of collected questionnaires is 51. Although the number and composition of participants is inappropriate for drawing rigorous scientific conclusions, we already got interesting initial results and gained valuable experience to carrying on with larger-scale studies.

3.1 A: Inter-Personal Differences

Table 3 shows the minimal and maximal vectors observed in the experiment, and distances between them. It is clear that respondents' answers cover a large proportion of decision space between the possible extremes (which are all 1's and all 4's), and that the maximum distances are large, particularly with *Store*.

Table 3: Minimal and maximal observed vectors.

Task:	*Car*	*Store*
Min:	⟨111111221124⟩	⟨1112111111121111⟩
Max:	⟨123423443344⟩	⟨3444342434343444⟩
Distance:	0.53	0.81

Observing individual vectors also reveals a great variability of answers. In 51 questionnaires, there are as many as 47 different vectors for *Car*. This means that most of the vectors are distinct. There is a single most frequent vector ⟨111212341234⟩, which appears only 3 times. *Store* is even more extreme with 50 different vectors and only one appearing twice.

Average distances between these vectors are shown in the *Inter-personal* column of Table 4. They were calculated on all pairs of vectors of the first trial (561 pairs). The differences are roughly 14% and 24% for *Car* and *Store*, respectively.

Table 4: Average distances between vectors.

Task	*Inter-personal*	*Intra-personal*
Car	0.141 ± 0.062	0.114 ± 0.068
Store	0.237 ± 0.090	0.146 ± 0.069

3.2 B: Intra-Personal Differences

Intra-personal differences were assessed on 17 pairs of questionnaires that were answered by the same participants at two different times. Average distances between vectors in this case are shown in the third column of Table 4. They are about 11% and 15% for *Car* and *Store*, respectively.

As expected, intra-personal differences are smaller than inter-personal ones. However, they are not *substantially* smaller: the ratios between inter- and intra-personal distances in about one month time were $\frac{0.114}{0.141} = 0.81$ for *Car* and $\frac{0.146}{0.237} = 0.61$ for *Store*.

These results indicate that subjects' preferences drift a lot over time and that it is difficult for individuals to provide the same decision rules twice. At this point, this study cannot really explain the drift; we can only speculate about the effects of changed preferences, changed decision context, bad memory and even imprecise, inaccurate or otherwise "elusive" nature of decision rules.

3.3 C: Consistency of Decision Rules

As expected, the majority (45 of 51 ≅ 88%) of *Car* decision tables defined by respondents contain consistent decision rules. Even in the remaining 6 tables, the level of consistency is high: one 92%, one 94% and four 97%. This indicates that for relatively small decision tables (*Car*) that use preferentially ordered criteria, the users are generally capable of formulating decision rules of perfect or very high consistency without any help.

The consistency of decision tables for *Store* is much lower, as shown in Figure 1. The average consistency is only 0.77. Again, this was expected to some extent, but not that much. In this study, we cannot really explain the reasons, which might be due to using preferentially unordered criteria, non-monotone consumer behaviour, or just due to the sheer size of the decision space to cover (four binary criteria). This remains an open challenge for future research.

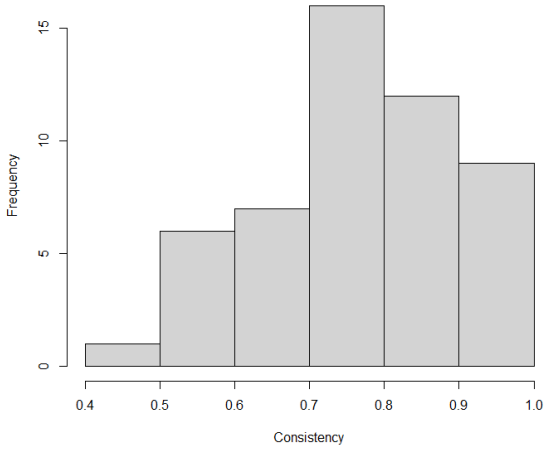

Figure 1: Consistency of *Store* decision rules.

Another interesting consistency-related question is: did those 17 participants that answered the questionnaire twice defined consistent (or, respectively, inconsistent) rules both times? For *Car*, the answer is no. None of the participants defined inconsistent rules both times; there were 3 participants who improved and 2 participants who degraded their consistency with time. With *Store*, most of the participants formulated inconsistent decision tables; 6 of them improved and 10 degraded the consistency measure.

3.4 D: Assessment of Weights

Figure 2 displays a boxplot of weights of *Car* decision rules as assessed by the participant (Ω) and by the three methods defined in section 2.3: LA, IG and GG.

On the one hand, we can see that participants, in average, assessed the two criteria, *Price* and *Technical characteristics*, almost equally, with a slight statistical leaning towards the latter. On the other hand, all the methods clearly indicated that, according to defined decision rules, *Technical characteristics* are far more important than *Price*. Here we can claim that users did not assess their weights really well, while the methods were largely consistent with each other.

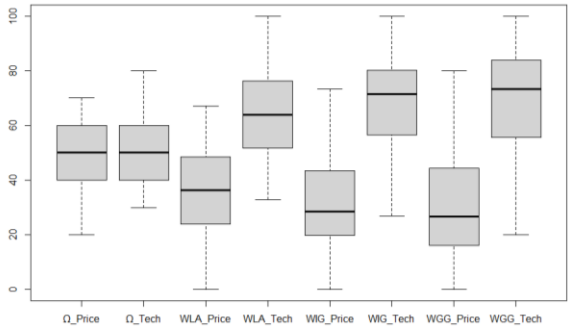

Figure 2: Weights of *Car* criteria assessed by the participant (Ω) and three methods, LA, IG and GG.

Results for *Store* are shown in Figure 3, where participants' weights (Ω) are compared with those assessed by the IG method. In contrast with *Car*, at least the order of criteria's importance was estimated correctly by participants (*Size* being the least and *Quality* the most important). Again, comparing human and algorithmic weights, the former are less extreme and all leaning towards 20–30%, while the latter are more extreme, ranging from about 10% to 50%. Yet again, the three methods turned out similar to each other (these results are not shown here).

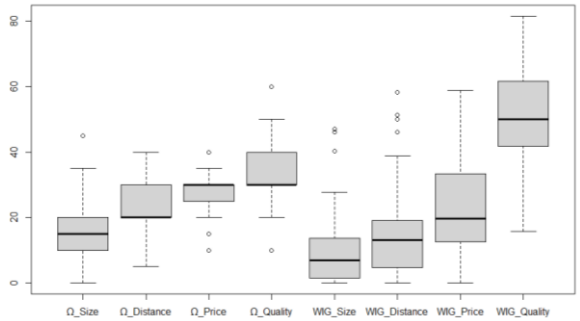

Figure 3: Weights of *Store* criteria assessed by the participant (Ω) and method IG.

Table 5: Differences between weight assessments of participants (Ω) and methods LA, GI and GG.

	Car			*Store*		
	LA	GI	GG	LA	GI	GG
Ω	0.086	0.137	0.159	0.062	0.099	0.116
LA		0.065	0.087		0.063	0.080
GI			0.038			0.034

Table 6: Weight differences between two trials (*Car*).

Car	Ω	LA	GI	GG
Ω	0.089	0.154	0.159	0.192
LA		0.127	0.114	0.152
GI			0.102	0.110
GG				0.126

Table 7: Weight differences between two trials (*Store*).

Store	Ω	LA	GI	GG
Ω	0.080	0.112	0.128	0.134
LA		0.098	0.121	0.121
GI			0.116	0.124
GG				0.126

The results of comparing all pairs of weight assessments (by Ω, LA, GI and GG) and calculating their difference using the formula from section 2.4, are shown in Table 5. Generally, differences are small, and particularly small differences (in the 4% range) are between GI and GG. Differences between LA vs. GI and GG are somewhat larger (about 7%). Differences between Ω and the remaining three methods are all greater than 6%. However, coming as a surprise and contrary to our expectations, the differences between Ω and LA turned out substantially lower than those of Ω vs. GI and GG. This indicates that the method LA, as implemented in DEXi, actually relatively well resembles user-assessed weights even for non-monotone decision tables.

Finally, Tables 6 (for *Car*) and 7 (for *Store*) compare weights between the first and second trials for the 17 participants that answered the questionnaire twice. The most important observation is that weights do change in time, however, the participants' assessments change less than those of the three methods. Roughly, we may expect about 9–19% weights change between subsequent trials, which are separated by about one month time.

4 Conclusion

At some point, we began calling this study *Aurora*. Considering the discrete nature of DEX decision tables, one might assume that there is little freedom in defining "vectors" of decision rules and that all decision tables should look the same. As we obtained more results from the study, it became increasingly evident that the opposite is in fact true. Humans are incredibly diverse and almost all decision tables, obtained in this experiment, were distinct. The search for "accurate" and "stable" preferences, represented with decision rules, turned to something like a quest for *Aurora Borealis*, which is real, but constantly changing its shape and colours.

The most important findings of this study are:

- Decision makers' preferences, expressed in terms of decision tables, are very diverse, peaking at inter-personal differences of 53% and 81% in the two investigated tasks (*Car* and *Store*), and the respective averages of 14% and 24%.
- Preferences of a single decision maker in time (intra-personal differences) also change a lot. Not as much as inter-personal differences, but close (11% and 15% in this study).
- For relatively small tables and preferentially ordered criteria (such as the *Car* task in this study), decision makers are perfectly capable of formulating consistent (or almost consistent) decision tables without any instructions and any supporting software tools.
- For larger decision tables with less clear ordering of criteria values, the consistency deteriorates

drastically. The reasons are yet to be determined by additional studies.

- Not only do the decision tables change over time, but the criteria weights, which are provided by the user and assessed from decision rules, also vary. The observed differences in criteria weights range between 9% and 19%.

Due to the small number and possibly biased composition of participants, these results should be considered preliminary. However, they have already provided some indications that contradicted our initial expectations, such as those that DEX decision tables are "rigid" and the variation of decision rules is small.

In the future, we wish to carry out a similar study on a larger and more diverse sample of participants. By doing so, we can establish concrete research conclusions and compare them with experience using common numeric weight-based MCDM methods.

The main challenge remains the question of how to "catch" participants, who are by default anonymous, for their second trial in a not too random time interval between the trials. Also, we wish to avoid respondents' "cheating", i.e., remembering and submitting exactly the same answers in both trials. Building on the results of this study, we aim to develop methods for identifying reasons for decision-rules drifting over time and their potential inconsistency. However, as we are pleased with the simplicity of the current questionnaire, we wish not to complicate it further with additional research questions.

Acknowledgments

The author acknowledges the financial support from the Slovenian Research Agency, research core funding P2-0103. The author also expresses his gratitude to students of the University of Nova Gorica, colleagues from Jožef Stefan Institute and friends who participated in the study.

References

Bohanec, M., & Zupan, B. (2004). A function-decomposition method for development of hierarchical multi-attribute decision models, *Decision Support Systems* 36, 215–233. doi: 10.1016/S0167-9236(02)00148-3

Bohanec, M (2020). *DEXi: Program for Multi-Attribute Decision Making, User's Manual, Version 5.04*. IJS Report DP-13100, Jožef Stefan Institute, Ljubljana, 2020. https://kt.ijs.si/MarkoBohanec/pub/DEXiManual504.pdf

Bohanec, M. (2022). DEX (Decision EXpert): A Qualitative Hierarchical Multi-criteria Method. In: A. J. Kulkarni (Ed.), *Multiple Criteria Decision Making* (Vol. 407, pp. 39–78). Singapore: Springer. doi: 10.1007/978-981-16-7414-3_3

Deguine, J.-P., Robin, M.-H., Camilo Corrales, D., Vedy-Zecchini, M.-A., Doizy, A., Chiroleu, F., Quesnel, G., Païtard, I., Bohanec, M., & Aubertot, J.-N. (2021). Qualitative modeling of fruit fly injuries on chayote in Réunion: Development and transfer to users. *Crop Protection* 139, doi: 10.1016/j.cropro.2020.105367.

Greco, S., Ehrgott, M., & Figueira, J. (2016). *Multi Criteria Decision Analysis: State of the art Surveys*. New York: Springer. doi: 10.1007/978-1-4939-3094-4

Kulkarni, A.J. (2022). *Multiple Criteria Decision Making*. Studies in Systems, Decision and Control 407, Singapore: Springer, doi: 10.1007/978-981-16-7414-3_3.

Raileanu, L.E., & Stoffel, K. (2004). Theoretical Comparison between the Gini Index and Information Gain Criteria. *Annals of Mathematics and Artificial Intelligence* 41(1):77-93. doi: 10.1023/B:AMAI.0000018580.96245.c6

Rezaei, J., Arab, A., & Mehregan, M. (2021). Equalizing bias in eliciting attribute weights in multiattribute decision-making: experimental research. Journal of Behavioral Decision Making, 35(2). doi: 10.1002/bdm.2262

Rokach, L., & Maimon, O. (2015). *Data Mining with Decision Trees: Theory and Applications*. New Jersey: World Scientific. doi: 10.1142/9789812771728_0001

Saaty, T.L., & Vargas, L.G. (2012). *Models, Methods, Concepts & Applications of the Analytic Hierarchy Process*. New York: Springer. doi: 10.1007/978-1-4614-3597-6

Silva, F.F., Souza, C.L.M., Silva, F.F. et al. (2021). Elicitation of criteria weights for multicriteria models: Bibliometrics, typologies, characteristics and applications, *Brazilian Journal of Operations & Production Management* 18(4). doi: 10.14488/BJOPM.2021.014

Thakkar, J.J. (2021). *Multi-Criteria Decision Making*. Studies in Systems, Decision and Control, Vol. 336. Singapore: Springer. doi: 10.1007/978-981-33-4745-8

Vetschera, R., Weitzl, W., & Wolfsteiner, E. (2014). Implausible alternatives in eliciting multi-attribute value functions. *European Journal of Operational Research*, 234(1), 221–230. doi: 10.1016/j.ejor.2013.09.016