QUANTITATIVE SUPPORT FOR UX METHODS IDENTIFICATION: HOW CAN MULTIPLE CRITERIA DECISION MAKING HELP?

Paulo Melo[1]

Faculty of Economics, University of Coimbra, 3004-512 Coimbra, Portugal

INESC Coimbra, Rua Antero de Quental 199, 3000-033 Coimbra, Portugal

Luísa Jorge

Polytechnic Institute of Bragança, 5301-857 Bragança, Portugal

INESC Coimbra, Rua Antero de Quental 199, 3000-033 Coimbra, Portugal

Abstract

In this paper we present views on how quantitative Multiple Criteria Decision Making (MCDM) approaches may be applied to certain aspects of User Experience (UX) Design and Evaluation (D&E) methods identification, emphasizing its strengths and weaknesses for this task. Often D&E methods need to be applied in contexts different of those they had been applied before, and as such must be transferred to those new contexts. In this work we present a model for the quantitative method matching step of the transfer process, describe how different MCDM methods can be applied to this task and discuss the results of an experience that tried to apply a couple of MCDM methods to method selection.

Keywords: Multicriteria matching; MCDM Matching; Usability Evaluation Methods Identification; PROMETHEE; SMAA-2

1 INTRODUCTION

To support the task of usability and user experience design and evaluation, several procedures, called usability (or sometimes, user experience) evaluation methods, have been proposed through the years (E. L. Law, Springett, and Winckler 2009). However, not all methods are well suited or may be used in all situations. In general, methods present a set of characteristics, which may make them better matched for particular situations or problems (usually those to which they were proposed to solve). In this work we intend to show how MCDM could be applied to help perform this selection using the knowledge of those characteristics.

[1] Corresponding author: pmelo@fe.uc.pt

Notice that while it can be argued that a method is not necessarily the adequate unit to be used when describing evaluation transfer (Woolrych et al. 2011), the same considerations could be extended to smaller or larger units (e.g. ingredients or meals in the work cited, although in this work we will use just the term "method" to describe a component of an design or evaluation process with reproducible characteristics that can be applied to different problems).

A few additional scenarios may also require finding a particular method for a particular situation:

- You know what kind of characteristics you require on your UX evaluation, but don't know which methods will provide them.
- You want to learn about methods sharing characteristics with a "favorite" UX method.
- You want to find the methods compatible with universal access, or create universal design resources to allow the selection of such methods (see, e.g. C. M. Law et al. 2007)

To find which methods are best suited to particular UX usage circumstances or usage situation (henceforth just called "situation"), we will need to match the known characteristics of the methods to the requirements of the situation (and likewise, the known necessities of the methods to what the method user is able to provide in that situation). If a method is being applied in a different application domain (so called method transfer), domain-specific information may need to be made explicit as situation characteristics and method characteristics (e.g., some characteristics may be present in a domain but not in another, so a method that requires such characteristics would not be able to be applied). In this work it is assumed that the methods will not see their characteristics changed by adaptation during the process of application to another domain (or, at minimum, the characteristics that are considered relevant for the choice problem will not be changed by the process).

Besides the "characteristics immutability" prerequisite described, the procedures discussed in this article also require that the information regarding method and situation characteristics are knowable, at some level (or at least that a subset of those is knowable, if procedures handling incomplete information are used). That is, at the level we require the "method" identification to happen (be it as "methods", "patterns", "ingredients" or "recipes", see Woolrych et al. 2011) its immutable characteristics (both in what it what it provides to its practitioners and what requires to be used) must be able to be described. An additional prerequisite of the procedure is that the characteristics of the situation to which the method is to be applied are also know (both in what is required of the method in that situation – situation requirements – and of which resources can be provided). In an ideal situation we would then proceed to match the mutual requirements to the mutual resources, to find the "best match". In this work it will be shown how Multi Criteria Decision Making

(MCDM) techniques could be applied to this task, and through an example depict some results and uncover a few practical problems in using such techniques.

In the next section a model for quantitative method matching is proposed (section 2), followed by the description of a practical exercise using this model. The results of this exercise are then analyzed. Next, alternative approaches are discussed, and we conclude by commenting the approach strengths and weaknesses.

2 A MODEL FOR QUANTITATIVE METHOD MATCHING

In this section we will try to describe how a quantitative model can be used to support the task of matching methods to situation requirements. Notice that for this model to be used there is a need for at least two kinds of actors: those who provide the knowledge about the methods (supposedly constant and therefore applicable to different situations) and those who provide the information regarding the situation requirements. In an actual MCDM environment, these latter actors, which will be called Decision Makers (DM), will also be required to provide information regarding their preferences concerning method characteristics, so the procedure can select among UX methods that are incomparable (if a method is best suited according to a certain characteristic while other is the best suited according to another, preference information regarding characteristics is required to select among them). According to the MCDM procedure used, the presence of a third kind of actors, facilitators or DM analysts, may also be advisable (although its presence may be lessened by the usage of simple procedure and/or help by automatic tools).

2.1 Problem Formulation

The problem to solve in short is: how to find the (set of) method(s) best suited to a particular situation, from a collection of methods?

To solve this problem there are a few required inputs, which should be provided in the adequate detail:

- Situation Description: definition of the characteristics of the situation, namely those that influence the choice of methods. The characteristics need to include what the methods should provide (requirements) but also the resources (time, money, knowledge) that are available.
- Set of Methods: the set of UX D&E methods users are willing to consider for the particular situation. Notice that those methods are assumed to be mostly static, that is either their characteristics are assumed as not being altered by the transfer process or the characteristics of the method resulting from the transfer process are assumed to be fully known. Only such methods will be considered as

"potentially transferable" using this procedure. For each method information must be collected regarding what its application provides and which resources are essential for it to be used.

To operationalize the choice using MCDM methods, both the situation description and the characteristics regarding the methods must be known in a complete and objective way (one which is shared by all the participants in the transfer process). There must be a synergistic description between situation and methods, in that the characteristics that are presented as requirements in one should possibly be present as resources in the other, at some satisfaction level (it is possible to have resources that don't map to needs, but the decision procedure will not take those into consideration). If there is uncorrectable lack of details or lack of shared understanding regarding any particular characteristic, that will make that characteristic unsuitable for quantitative manipulation, and at worst may prevent it to be used by the decision methods.

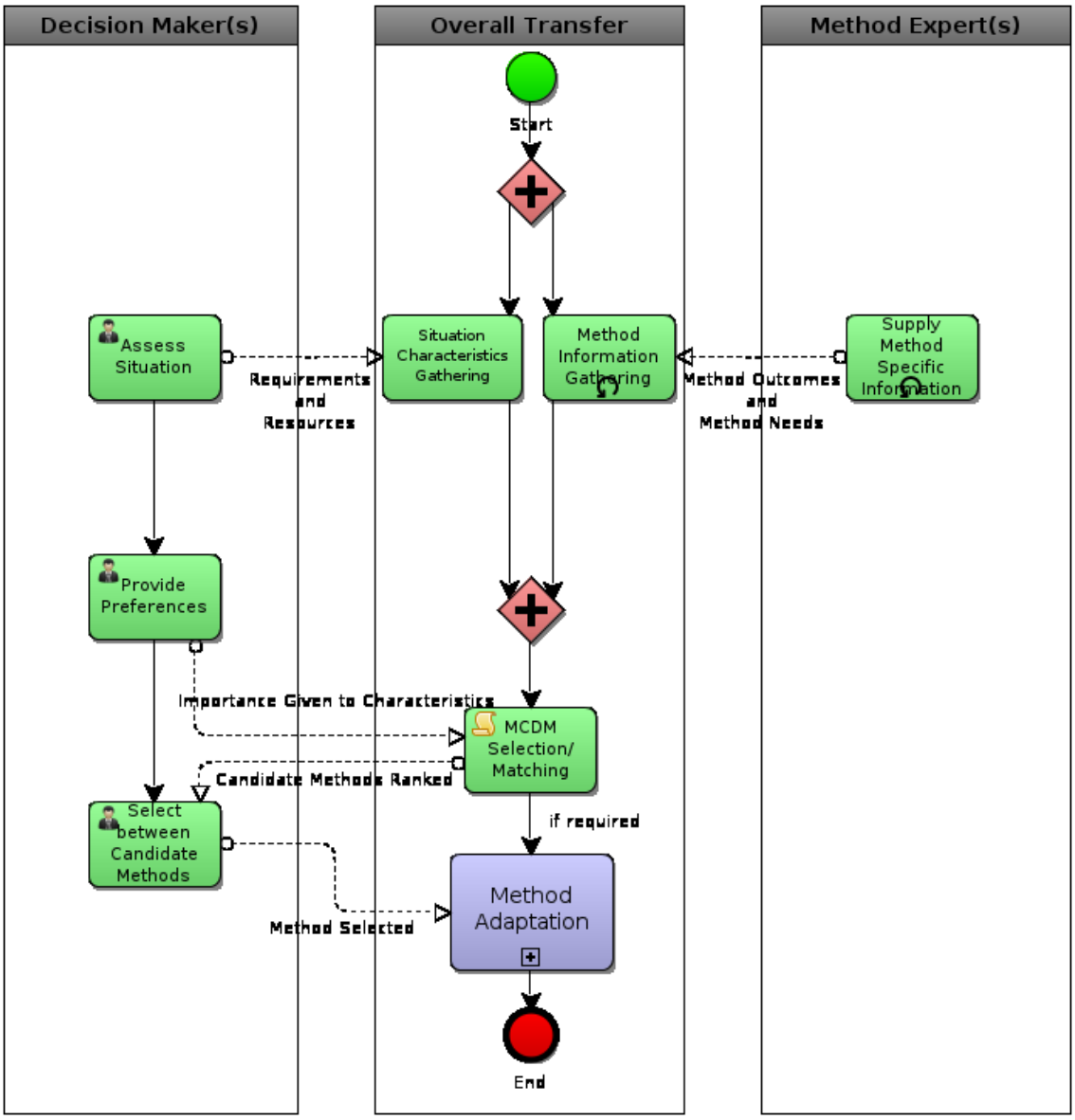

Figure 1 – Overall process structure

The overall structure for transfer proposed is depicted in Figure 1. As it can be seen, the process requires the participation of two (sets of) actors, one providing situation specific knowledge and preferences (the decision maker) and other(s) providing information regarding method characteristics in terms of their outcomes and requirements for application (termed method expert(s)). The outcome of the process is a method suitable for application in that situation, which may require domain and situation specific adaptations (this stage although depicted is not explored in this article). As can be seen in the figure, previous to the actual method selection and adaptation, several information gathering and processing (MCDM selection/matching) stages must be performed. In the following sections these will be described in greater detail, using the nomenclature proposed below.

For a practitioner, following the process presented in Figure 1 could provide a series of task steps to help discover which method could be applied in a particular situation, in the role of the "Decision Maker":

- assessing the situation s/he would start in initial "discovery and gathering" steps, where s/he would describe the characteristics of the particular problem situation s/he is trying to support in terms of situation requirements and resources that can be provided to the method (that description would necessarily need to be provided in terms that could be matched to method needs and method outcomes delivered by UX experts with method specific knowledge – method experts);
- the relative importance of the characteristics described previously would then be ranked by the practitioner, according to her/his preferences and to the interaction requirements of the MCDM procedure used;
- the MCDM procedure would be followed and return a ranking of candidate UX D&E methods (a ranking that can include incomparability between some methods, as a result of the previous inputs);
- finally the practitioner would then be required to select among the ranked candidate methods the one(s) s/he would deliver to a final "Method Adaptation" step, which could be more or less extensive in function of method and situation characteristics (since the method would be selected mostly in terms of fit to situation, but that fit would not necessarily be perfect).

2.2 Nomenclature

Let us assume that we have a set $M = \{m_i\}$ of (potentially transferable) design and evaluation methods. From the characteristics, we need to derive a set *C* of criteria (method characteristics). For each criterion/method characteristic *k*, each method will have an evaluation $g_k(m_i)$. From the situation description we also obtain a problem description *P*, which has its own evaluation on each criterion, $g_k(P)$.

While the derivation of the criteria and evaluations from the characteristics is not a trivial process, as seen in (E. L. Law, Springett, and Winckler 2009) and as our own experiment described in this text will show, let us assume that we have such description. The problem of finding the correct method becomes thus to find the method or set of methods "most similar" to the problem description according to the criteria set.

2.3 Simple (naïve) formulation

To find the method "most similar" to the problem description, one needs to find a way to measure the similarity. Let's define $A_P^{m_i}$ as a measure of the "Distance" between the problem *P* and the characterization of *method i* (m_i).

According to each criterion *k*, we will have a measure of this distance considering that particular criterion, which we will denote as $h_k\left(A_P^{m_i}\right)$.

Several kinds of measures can be proposed as $h_k\left(A_P^{m_i}\right)$. Here we present two:

A. Let's assume that the criteria chosen are distinctive enough that just its presence or absence in the problem description is enough to include or exclude a method from use (that is, each requirement must unavoidably be met by the corresponding resource). In this case, we can use:

$$h_k\left(A_P^{m_i}\right) = \begin{cases} 1 & \text{if the criterion } k \text{ is present both in } P \text{ and in } m_i \\ M & \text{otherwise} \end{cases}$$

M would be defined as an integer value bigger than any other value in computation. A direct implementation of this formulation may lead to numerical problems since it requires computation with large numbers, therefore it can be stated that if *M* appears in the evaluation of a method m_i, such method should be excluded from further consideration in the decision process.

It should be noticed that this model is generic enough to encompass common mechanisms of method selection used in practice:

- If one just wants to present all the methods that include a particular set of characteristics, $h_k\left(A_P^{m_i}\right)$ will be defined precisely as described.
- If we try to rank the "best" methods that share a set of characteristics, like is in practice done in (Ferre Grau and Bevan 2011), then the previous approach should be slightly altered (again, the use of *M* can be replaced by actual exclusion of methods from consideration):

$$h_k\left(A_P^{m_i}\right) = \begin{cases} g_k(m_i) & \text{if the criterion } k \text{ is present in } P \\ M & \text{otherwise} \end{cases}$$

This simple mechanism, however doesn't take into account the fact that the degree of satisfaction of a characteristic may not be "all or nothing", that is, that some method characteristics may be only partially required for a particular situation (or required at different levels in different situations), and that methods may differ in their degree of support for different characteristics.

Therefore, we propose also a more generic formulation:

B. We can try to compute the difference among the evaluations of the problem and the method according to criteria *k*:

$$h_k\left(A_P^{m_i}\right) = |g_k(m_i) - g_k(P)|$$

This formulation assumes that $g_k(X)$ can be computed accurately for all methods and all problems descriptions on all criteria. It further assumes $g_k(X)$ to be numerical/cardinal and that the results from different criteria share the same scale factor (though both assumptions can be relaxed by redefinition of the norm to be used). This formulation can be further improved (at cost of additional complexity) by supporting the exclusion of methods assumed to be sufficiently different from *P* (although such exclusion can also be made in the subsequent process, using suitable MCDM procedures supporting the concept of veto (Figueira, Mousseau, and Roy 2005)).

This is a simple formulation, and can be used with any kind of MCDM procedure. It is therefore possible of application for any transfer process for which the information can be gathered. The information required (evaluation on criteria of methods and situation) is an area where work has been developed, e.g. by MAUSE Working Group 2 (E. L. Law, Springett, and Winckler 2009). It should be noticed that the process can be used only to choose among known "methods" (combining methods or determining which methods to combine is not supported by this formulation, although composite methods/meals whose properties are known could be supported as choices).

2.4 Resolution using simple additive weighting

Once the distance between a method m_i and *P* can be expressed for each criterion as $h_k(A_P^{m_i})$, different MCDM procedures could be followed to determine the actual method to be used. One of the simplest MCDM procedures would be to apply a SMART-like (Edwards 1977) approach using simple additive weighting.

If we elicit a set of weights $W = \{w_k\}$, where w_k reflects the importance given to method characteristic *k* for the process. In this text, while we assume that the decision is being made by a group, these weights are considered to be shared by the group, and so can be elicited using any group tool. It should be noticed that the weights are assumed as independent of the actual transfer process being done (i.e. are not considered dependent on the situation *P*), although different sets of *W* for different classes of transfer practices could be envisaged.

$$\text{cost}_{\text{m}_i} = \sum_{k \in C} w_k \, h_k(A_p^{m_i})$$

Using these weights, we can try to find the method j with lowest cost ($\text{m}_j : \min_{\text{m}_i \in M} cost_{\text{m}_i}$) simply by ordering the costs.

This method is simple and easy to implement, but has a few difficulties for practical adoption. It assumes that the evaluation function $h_k(\ \ )$ is cardinal (which can be trivially proven to be false for some simple functions,

even the first one proposed, if care is not taken on building the characteristics). However, this difficulty can be lessened by judicious choice of evaluation functions.

It further assumes full comparability and a compensatory model (implied by simple addictive weighting), which means that large differences between a particular method (m_i).and P on a characteristic (that is, a method that is really not well suited for a particular situation, on one characteristic) could be overcome by having small differences on the remaining. Whether a compensatory model can be used will depend on the actual characteristics chosen and on the strictness of the matching required.

3 APPLYING THE MODEL

To verify whether the simple formulation described above could be applied in practice, we performed a practical experience to discover whether a group could cope with the informational requirements to apply MCDM on a method selection task. This experiment was performed by a set of 20 UX experts, organized in four groups (O1 to O4). The exercise intended to provide an answer to which method would be "best suited" to a particular situation.

3.1 Application Context

In the actual exercise, we restricted our choice to six methods, chosen to provide a sufficiently diverse mix of characteristics:

- Heuristic Evaluation (Nielsen and Molich 1990)
- Cognitive Walkthrough (Nielsen 1994)
- UX curve (Kujala et al. 2011)
- Eye tracking (Yarbus 1967)
- Card Sorting (Nielsen and Sano 1995)
- Contextual Inquiry (Wixon, Holtzblatt, and Knox 1990)

And tried to select which one best suited this particular situation:

- As a serious game designer I want to use collaboration techniques from e-learning applications to improve the effectiveness of my developed games.

The situation was deliberately stated in an open formulation, so that there could not be an "a priory" obvious solution to the group, and members could discuss among themselves what characteristics would be required in this kind of situation.

To assert fit between method and situation, the experiment was restricted to six method attributes/criteria (notice these criteria are shown as requirements for the method, but they could also be stated as resources available in the situation –these criteria were selected from a shortlist, provided by UX experts, of relevant method characteristics, but other criteria could be used instead):

- *Depth*: Does the method go in-depth in its results (e.g., very detailed usability problems), or does it provide just some general hints or ideas?
- *Documentation Level*: Documentation of procedure, is it a method that is well documented and prescribes actions in much detail, like, for instance, GOMS (Card, Moran, and Newell 1986), or is it mainly a general approach (e.g., usability testing)?
- *Structure*: Level of structure/degree of formality. Some methods, e.g., GOMS are very structured and formal; Heuristic Evaluation is rather unstructured and informal.
- *Time*: How much time does it take to perform the method to its completion?
- *Expertise*: How much expertise is required to use the method? Some methods require a certain level of expertise in some domain (e.g., usability or application domain).
- *Phase*: When is the method best applied? Some methods are best suited for early design phases, other are for final products only, and other still for inspiration phases, etc.

Table 1– Scales used for Method Characteristics/Criteria

Method Characteristic	Scale to Use	Direction
Depth	Likert 5 points	
Documentation Level	Likert 5 points	
Structure	Very informal, Informal, Formal, Very formal	
Time	30 seconds to several years	Minimize
Expertise	Novice, Familiar, Knowledgeable, Expert	
Phase	Design concept, Mockup/Paper, Functional prototype, Functioning product	

As a previous step, and to familiarize themselves with the process, the groups were asked to create evaluation scales (in fact, define the evaluation domain) for each of the six characteristics. The results are presented in Table 1. As can be seen, there was a need to use Likert-like items for some criteria, since the groups could not agree on common descriptions to the method characteristics (for the applicability of Likert scales to similar applications, see Norman 2010). Notice also that for most characteristics the groups could not agree on a "desirability" direction associated with the characteristic, which means that using it directly as a criterion (e.g. using the measure A described previously) would be difficult.

3.2 From collected data …

The four groups evaluated both the methods and the situation according to the previous method attributes (in such a way that each method attribute would be evaluated by two different groups – this was made to both lessen the load on each individual group and also to mimic a decentralized application where not everyone knows all the methods or its characteristics). Table 2 presents the data received from the 4 groups. The raw data of choices made by the different groups was aggregated (e.g. when group elements defined several values for a particular criterion/method attribute, an average group value was used). Time values were converted to a common unit. This data was then organized in two "synthetic" groups, SG1 and SG2 each assembling some evaluations from three (of the four) original groups.

Converting the data in Table 2 to numerical inputs, by normalizing each value from each scale on Table 1 to a value in the 0-1 interval, gives the data in Table 3. Additional calculations, using Pearson correlations, has shown that the evaluations by the two groups on the different methods characteristics were usually similar (see Table 4), where 4 of the characteristics have correlation above 0.6. However, the two groups had a very distinct view on the situation *P*, which can be made visible by performing an additional Pearson correlation on the *situation index* column for SG1 and SG2 of Table 3. Doing so, we will obtain $\rho = -0.881$, indicating that there is a strong negative correlation between the groups on the two situation indexes.

Table 2 – Group Analysis by Criteria for Situation and Methods

		Criteria	*Situation Index*	Heuristic Evaluation	Cognitive Walkthrough	UX curve	Eye Tracking	Card Sorting	Contextual Inquiry
SG1	O1	Depth	*3*	1 (low)	4	2	5 (high)	3	3
	O1	Documentation Level	*4*	2	4	3	3	4	5 (high)
	O1	Structure	*v. informal*	v. informal	formal	formal	v.formal	informal	informal
	O2	Time	*28800*	300	300	9600	7200	480	7200
	O2	Expertise	*knowledgeable*	knowledgeable	knowledgeable	familiar	knowledgeable	familiar	knowledgeable
	O3	Phase	*functioning product*	functional prototype	functional prototype	functioning product	functional prototype	design concept	design concept
SG2	O3	Depth	*3*	4	4	2	2	3	5 (high)
	O4	Documentation Level	*3*	2	4	4	3	2	5 (high)
	O2	Structure	*formal*	v. informal	informal	formal	v.formal	formal	informal
	O4	Time	*2400*	480	480	2400	2400	240	4800
	O3	Expertise	*familiar*	knowledgeable	expert	novice	expert	familiar	knowledgeable
	O4	Phase	*design concept*	functional prototype	functional prototype	functioning product	functional prototype	mockup	functioning product

Table 3 – Group Analysis by Criteria for Situation and Methods (normalized)

	Criteria	*Situation Index*	Heuristic Evaluation	Cognitive Walkthrough	UX curve	Eye Tracking	Card Sorting	Contextual Inquiry
SG1	Depth	**0,500**	0,000	0,750	0,250	1,000	0,500	0,500
	Documentation Level	**0,750**	0,250	0,750	0,500	0,500	0,750	1,000
	Structure	**0,000**	0,000	0,667	0,667	1,000	0,333	0,333
	Time	**1,000**	0,002	0,002	0,328	0,244	0,008	0,244
	Expertise	**0,667**	0,667	0,667	0,333	0,667	0,333	0,667
	Phase	**1,000**	0,667	0,667	1,000	0,667	0,000	0,000
SG2	Depth	**0,500**	0,750	0,750	0,250	0,250	0,500	1,000
	Documentation Level	**0,500**	0,250	0,750	0,750	0,500	0,250	1,000
	Structure	**0,667**	0,000	0,333	0,667	1,000	0,667	0,333
	Time	**0,076**	0,008	0,008	0,076	0,076	0,000	0,160
	Expertise	**0,333**	0,667	1,000	0,000	1,000	0,333	0,667
	Phase	**0,000**	0,667	0,667	1,000	0,667	0,333	1,000

Table 4 – Correlation among Criteria between SG1 and SG2

Criteria	Pearson Correlation
Depth	-0,234
Documentation Level	0,630
Structure	0,818
Time	0,797
Expertise	0,883
Phase	0,325

Table 5 – Numerical Values for Distance between Method and Situation

	Criteria	Heuristic Evaluation	Cognitive Walkthrough	UX curve	Eye Tracking	Card Sorting	Contextual Inquiry	*StDev Dist*
SG1	Depth	0,500	0,250	0,250	0,500	0,000	0,000	*0,204*
	Documentation Level	0,500	0,000	0,125	0,250	0,250	0,250	*0,152*
	Structure	0,000	0,500	0,667	1,000	0,500	0,333	*0,304*
	Time	0,995	0,995	0,798	0,840	0,996	0,798	*0,092*
	Expertise	0,000	0,167	0,500	0,167	0,333	0,000	*0,178*
	Phase	0,333	0,333	0,000	0,333	1,000	1,000	*0,373*
SG2	Depth	0,250	0,250	0,250	0,250	0,000	0,500	*0,144*
	Documentation Level	0,250	0,250	0,125	0,000	0,000	0,500	*0,173*
	Structure	0,667	0,167	0,000	0,333	0,167	0,333	*0,208*
	Time	0,070	0,070	0,126	0,084	0,071	0,126	*0,025*
	Expertise	0,333	0,500	0,167	0,500	0,000	0,333	*0,178*
	Phase	0,667	0,667	1,000	0,667	0,333	1,000	*0,229*

3.3 … to MCDM inputs

As a result of the previous findings, it was decided to handle each group separately, but to average the evaluations of the methods on the characteristics where those evaluations converged (in fact, all but Depth and Phase). This way, the shared knowledge was used to improve the evaluation of the methods on Documentation Level, Structure, Time and Expertise, but separate group evaluations were used for the Situation description and for the Depth and Phase characteristics.

Using the procedure described previously, the difference on positions stated by each group on the different criteria and on the different situation descriptions was computed (using the measure B previously described), after converting those positions to a numerical value in the range [0,1] (see Table 5). With this calculation, a method that completely agrees with the situation on a particular criterion/method attribute would receive the value 0, whereas a method that is the opposite of the situation on that criterion would receive a value 1. Notice that the situation is seen as different among the two groups, and as such, even on "common" criteria, a method that is seen as "close" to the situation on one group can be seen as "away" from it by the other.

As we can see on Table 3 the evaluations diverged a little among the two groups. After applying the procedure described, from Table 5 can be seen that although the dispersion inside each evaluation (measured by the standard deviation on the values by criteria/group) is not very high, it should be noticed that even groups that achieve similar average distance to the situation on a method attribute don't usually select the same value for the evaluations (that is, although the average distance achieved is similar, the distance differ on the individual methods).

The most extreme case for different evaluations is found on the "time" criteria, where while both groups give not completely different evaluations on methods (see Table 2 for original values), they differ in such a marked way on the situation description that for a group almost all methods are close to the situation in terms of time while for the other almost all methods are very different from its situation in terms of time (as seen in Table 5).

An almost as stark difference among groups is found on the "phase" criteria, where a group considers the method should be applied on a functioning product where the other considers it should be applied on a design concept (although in this case, the results of the differences are not as visible as on the "time" criterion since the short ranges considered limit the divergence).

Notice that the algorithmic approach chosen to handle the differences is usually considered inadequate if the groups are to be considered as handling the precise same problem. A more common way to handle the divergences would be to reconvene the decision makers to try to get them to achieve a common

understanding of the situation (and of those criteria where marked differences were present). However, since this was impossible in the exercise conditions (and would also be very difficult to achieve in practical usage of this approach, where the "evaluators" of criteria and the "users" of the methods may never be in direct contact), the approach presented seems to satisfy the homogeneity of preferences requirement for MCDM procedure application

3.4 MCDM Procedures in Use

Once the data was computed according to the previous procedure, the actual MCDM process could be applied. For this exercise, two different MCDM methods were used, PROMETHEE (Brans, Vincke, and Mareschal 1986) and SMAA-2 (Lahdelma and Salminen 2001). Those MCDM methods where chosen because of their fit to the data collected, being also simple enough to not create particular difficulty in the results interpretation, while powerful enough to support scenarios and output incomparability (PROMETHEE) and imprecise/stochastic input data (SMAA-2).

The values were then input into the Visual PROMETHEE software (Mareschal and De Smet 2009), as shown in Figure 2. To support the different groups, the information was defined as two different scenarios, which will be compared (scenario 1 includes SG1 data, while scenario 2 uses SG2 data).

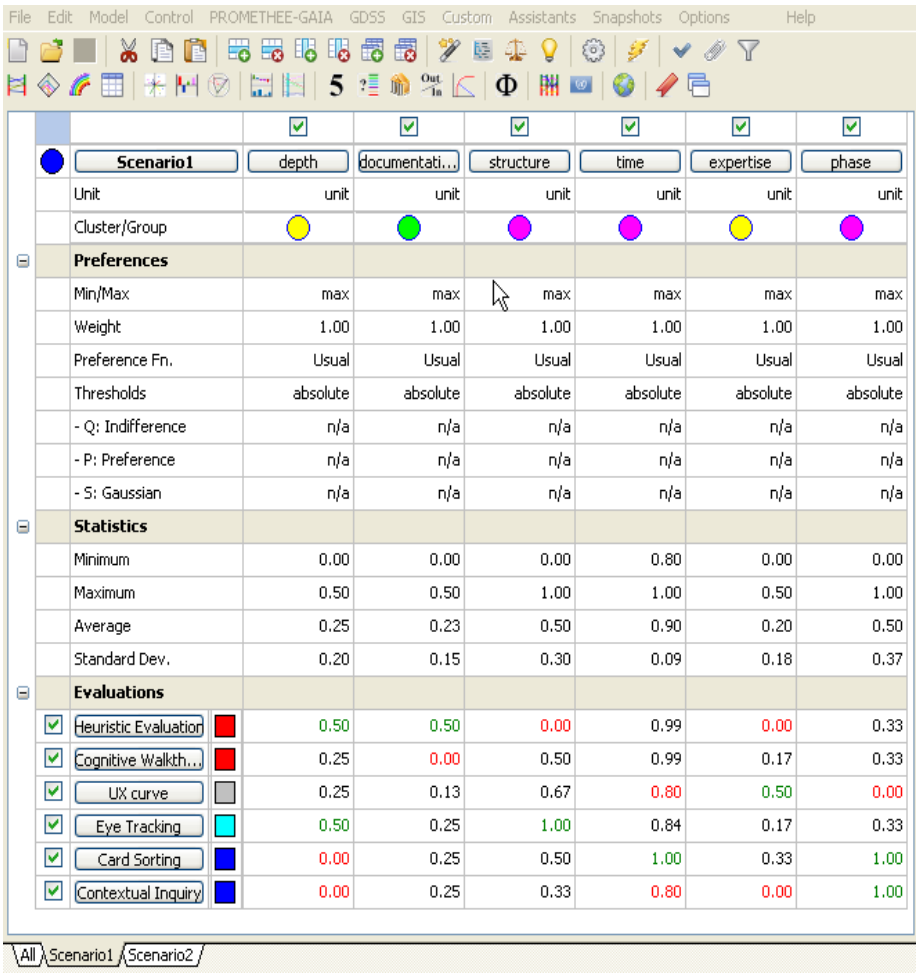

Figure 2 – Data used for PROMETHEE

With this data, we used PROMETHEE to assess the understanding from each group regarding which methods could be better applied to the situation. Assuming all method attributes are considered as equally important (equal weights on the method characteristics in Figure 2) PROMETHEE would give us the result present on Figure 3, presenting two partial rankings for the methods (the higher the methods are in the ranking the best the fit between method and situation). As can be easily seen the rankings are not the same, although they are in this case mostly in agreement (as measured by the PROMETHEE procedure, the Phi+ flow ranking relates to the number of comparisons where method A is seen as a better fit than the remaining methods, while the inverse of Phi- is related with the number of comparisons in which is seen as not being a worse fit than the alternatives (Brans, Vincke, and Mareschal 1986), so they don't represent precisely the same thing).

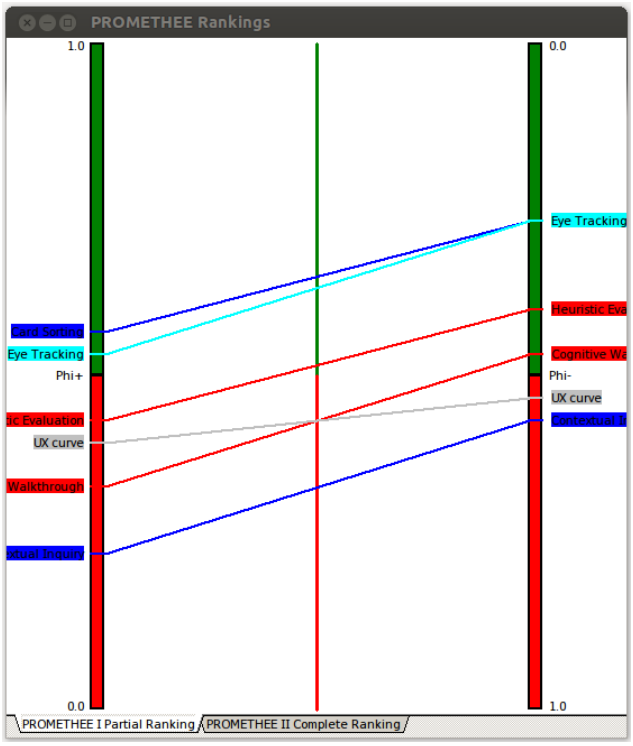

Figure 3 – PROMETHEE I partial rankings (SG1)

Although Figure 3 provides the full results of PROMETHEE, for non-initiates an easier interpretation of the results can be made when they are depicted as in Figure 4 (using the PROMETHEE Diamond, see Mareschal and De Smet 2009). In this picture is easy to understand that two methods are considered to be a better fit (Card Sorting and Eye Tracking) while the rest are seen collectively as somewhat similar among themselves and an overall worse fit for the situation (with Contextual Inquiry as the worse fit).

Although it could be claimed that given the results shown Card Sorting should be seen as a better fit than Eye Tracking, PROMETHEE by itself cannot give us an absolute ranking between the two, but can say that, giving equal importance to all method attributes, those two are certainly a better fit than the rest.

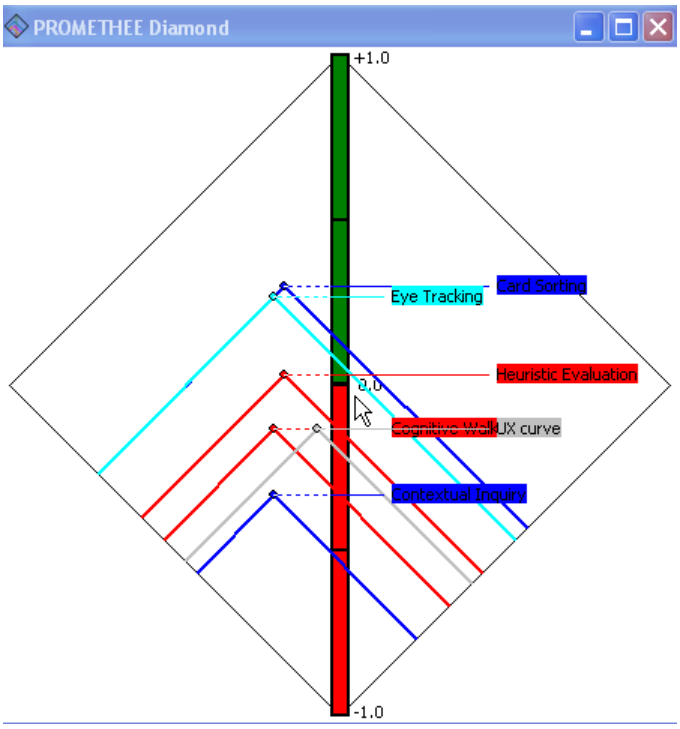

Figure 4 – PROMETHEE Diamond (SG1)

It should be noticed that the rankings presented are the combination of fit among different attributes, and there is some compensation among the different evaluations to reach the final rankings.

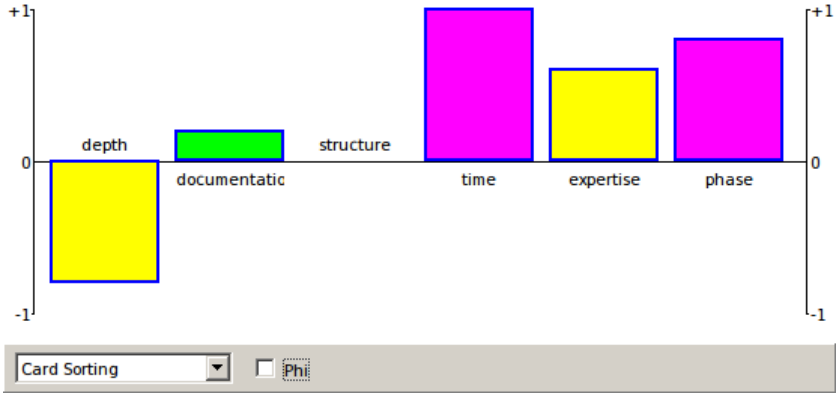

Figure 5 – Fit by method characteristic for Card Sorting (SG1)

As seen in Figure 5, Card Sorting is shown as having better fit on time, expertise and phase and worse fit on depth (with positive fit on documentation), whereas Eye Tracking (Figure 6) has an very good fit on structure and a little smaller fit on depth, and somewhat negative fits on time and phase. Overall those fits add up (using equal weights) to results that are broadly similar.

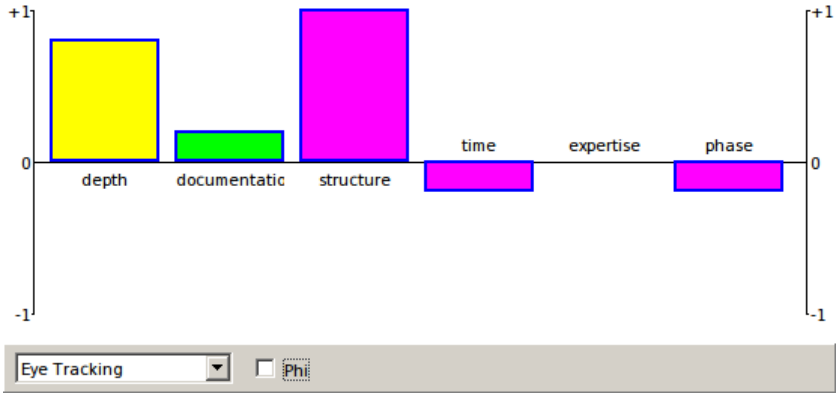

Figure 6 – Fit by method characteristic for Eye Tracking (SG1)

Some other methods (like Heuristic Evaluation, on Figure 7) can have worse fit on more method attributes and smaller fit with the remaining, which explains their relative ranking.

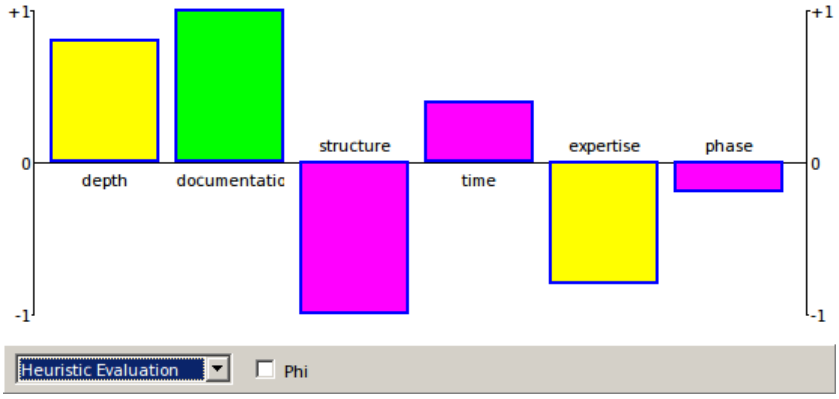

Figure 7 – Fit by method characteristic for Heuristic Evaluation (SG1)

Since the overall ranking is dependent on all the method attributes, a few good results may overcome extremely bad results on a particular method attribute. PROMETHEE supports the concept of veto to override this effect, but the experiment didn't use this possibility. Therefore, the results shown may overstate the fit of methods which would be disqualified by such veto.

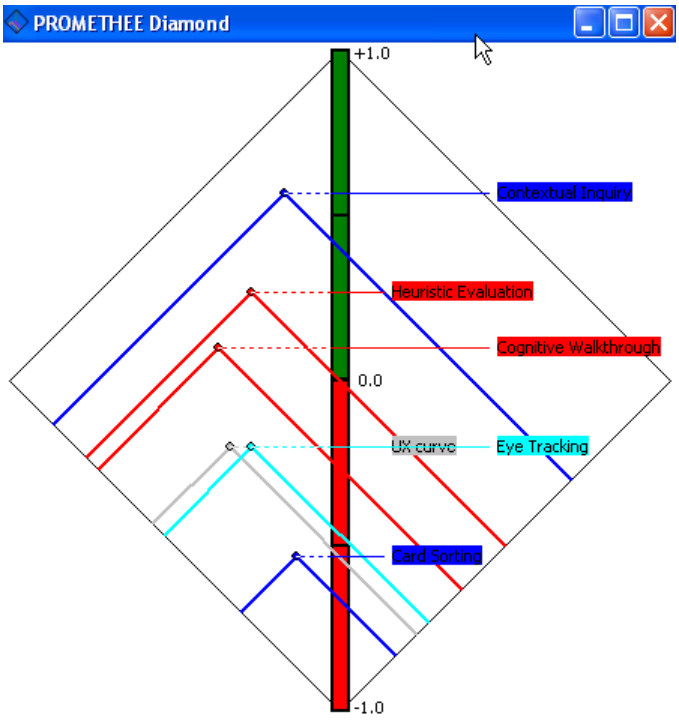

Figure 8 – PROMETHEE Diamond (SG2)

However, applying the same procedure as before to the data from SG2 will give us a very different ranking of methods (as seen in Figure 8). In fact even if we just use the PROMETHEE II ranking, it is easy to show that the two groups have completely different opinions regarding the method which is best suited for the situation (see Figure 9).

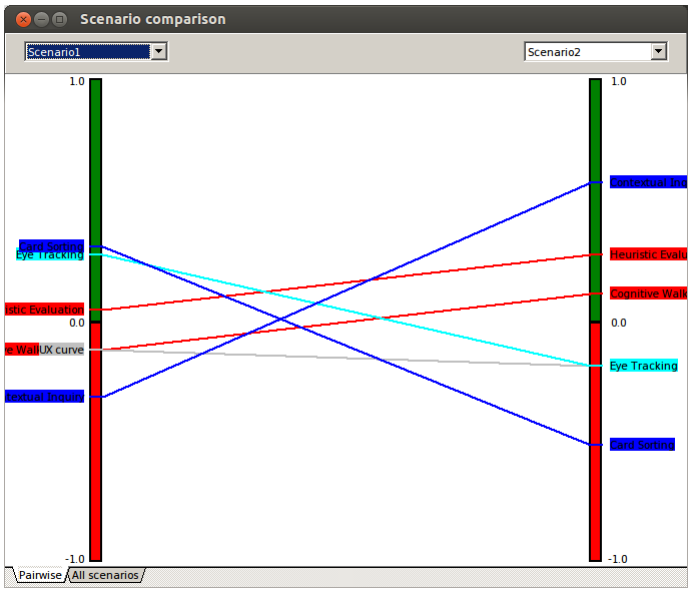

Figure 9 - Group rankings comparison

Notice however that these results mimic strongly the original negative correlation found in the situation description among the two groups. Since the situation perceived by those groups is mostly symmetrical, it isn't very surprising that a method that ranks best for a situation is ranked last in the other.

The results also depend on the relative importance given to method characteristics, but agreeing on importance to give each criteria can be hard. Some MCDM procedures can be used to provide a statistical distribution of results without fixing "a-priori" its importance (the MCDM procedures which do so are often called preference disaggregation approaches, see (Siskos, Grigoroudis, and Matsatsinis 2005)). This exercise was performed using the SMAA-2, the stochastic multi-criteria acceptability analysis procedure (Lahdelma and Salminen 2001). In this, each alternative is assigned a probability of belonging to a rank (from the 1, the best, to 6, the last one) in terms of weight combinations. As the results show (Figure 10) using the previous data, in almost all of the possible weight combinations "Card Sorting" would be selected as the best fit, and in over 95% "Contextual Inquiry" is the worst (although Eye Tracking is not always seen as the second best option). This shows a very strong agreement with the previous results, achieved using a different MCDA procedure and a different approach.

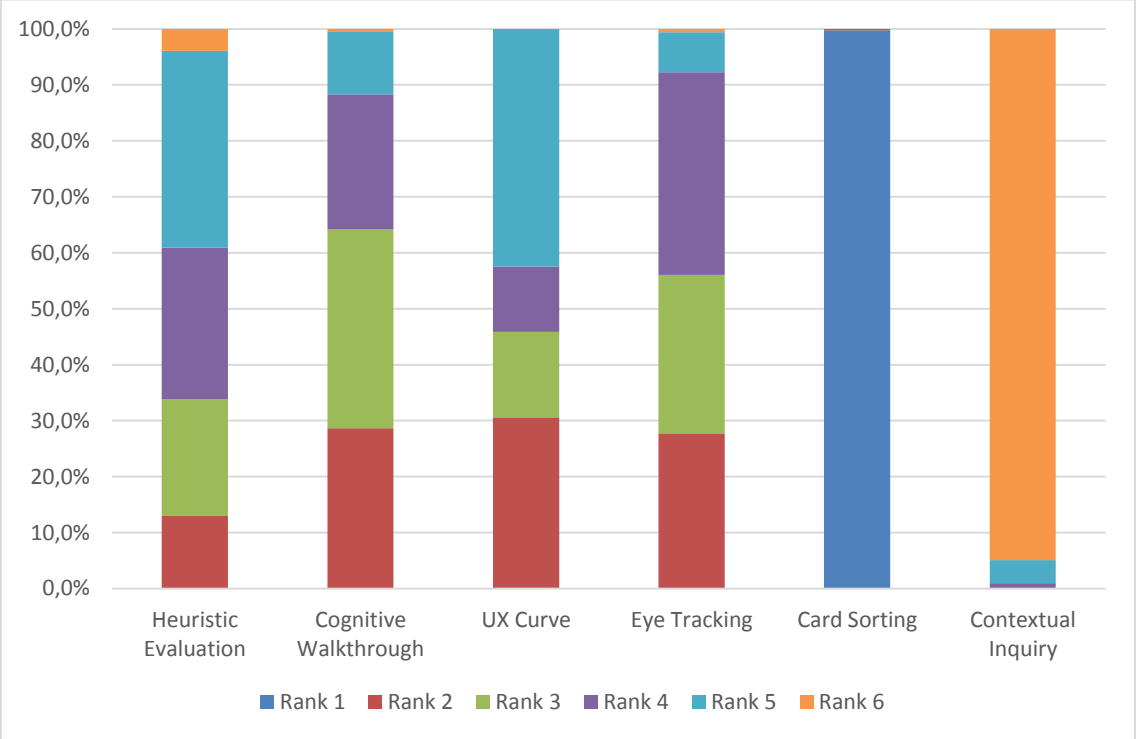

Figure 10 – SMAA rank acceptability indexes - using exact values (SG1)

Notice however that these results were arrived at using the average group evaluation for a few characteristics (achieved by averaging the evaluations of the pair of synthetic groups). To try to check whether such averaging distorted the results the experiment were remade in SMAA-2 using as parameters not exact values

but stochastic values with normal distributions with the same average and standard deviation that were achieved by averaging the groups. As one can see (Figure 11) Contextual Inquiry and Card Sorting are still overwhelmingly considered the worse and best fit, but there is now about 15% of evaluations where they don't have those ranks. But while some results don't agree so impressively with the previous ones, other are even more akin to the ones obtained with PROMETHEE, namely the classification of Eye Tracking. This reinforces the overall feeling of robustness for the results found.

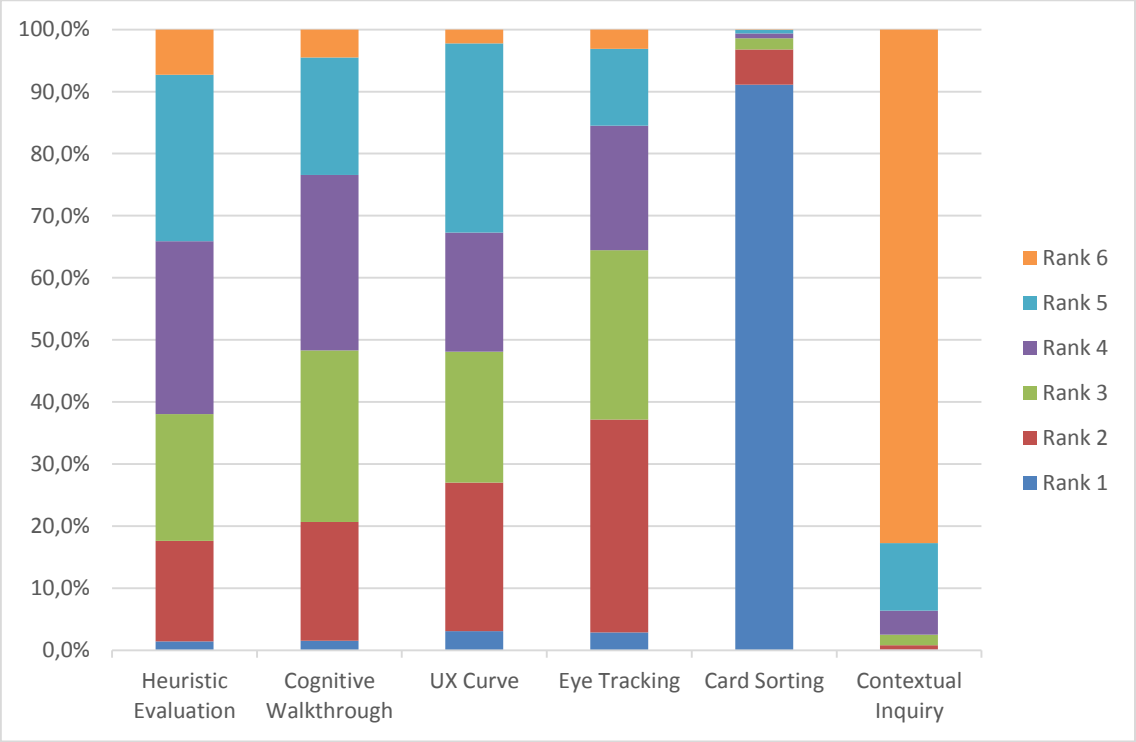

Figure 11 – SMAA rank acceptability indexes - using stochastic information (SG1)

4 RELATED WORK

The field of usability and user experience design and evaluation has generated through the years a significant number of methods and approaches (see, for instance, E. L. Law, Springett, and Winckler 2009 for a characterization of some of them). Their different characteristics mean that their application to all situations may not be desirable, or indeed even possible. As such, throughout the years, work has been developed to help with the task of selecting the adequate methods to particular situations, from books dedicated sets of methods (Nielsen and Molich 1990) to method characteristics reviews (E. L. Law, Springett, and Winckler 2009) to interactive tools to help method selection (Ferre Grau and Bevan 2011). However, in these approaches, the actual task of method selection is either left to the user, or is handled by a simplified approach, that doesn't provide the decision maker with a decision "audit trail" to justify the choice.

On the other hand, the selection from a set of alternatives possessing different characteristics of the most relevant subset is a long known problem addressed by multi-criteria decision methods, both in individual and group situations (see, e.g. Belton and Pictet 1997; Ehrgott and Gandibleux 2002). Those methods have been applied to a large number of problems, namely in computer science and service application matching (Amir et al. 2008; Medaglia and Fang 2003; Stroulia and Wang 2005), but the authors know of no previously published MCDM applications relating directly to the problem of matching UX evaluation methods.

The model proposed uses a simple distance measure to generate multiple criteria matching. Different scalarizing functions have been proposed in multi-criteria optimization (see, e.g. Miettinen and Makela 2002), and others could be proposed as adequate in this situation (namely to prevent, in the absence of veto, large differences in one criteria to overwhelm the comparisons). However, the choice of another scalarizing function by itself would not alter the process proposed (and could make the process less understandable for the decision makers if harder to compute, less intuitive metrics were used instead).

The limitations of using distance based approaches to matching problems should not be overlooked. However, while some alternatives have also been proposed to characteristics matching problems, like semantic matching (Ghomari and Ghomari 2009; Silva et al. 2010), which may be constructed removing some of these limitations, we consider that the multiple criteria approach proposed is both simpler and less dependent on the previous construction of a full domain ontology required to overcome the use of arithmetic distance definitions (although, as was stated, the requirements on information collection are not irrelevant even in the proposed approach).

5 COMMENTS ON THE APPROACH AND CONCLUSIONS

We presented a mechanism to allow for UX D&E method identification aid, using multiple criteria decision making tools. This mechanism can be used whenever there is a need to identify which method or methods fit the needs and capabilities available in a particular UX design or evaluation situation (namely when we try to describe whether methods previously applied on different domains should be considered to be applied on a different domain). The main contribution presented is therefore the model to support the identification of methods according to their fit to a particular situation, when both the method and situation characteristics are considered knowable. To verify applicability of this model, a practical experiment was performed and its results were analyzed.

This selection process may be performed by anyone/any group capable of describing the situation and the methods in terms suitable to the matching. This information on method and situation characteristics is the bare minimum required to apply the procedure. However, as it can be proposed that it would be hard for non-expert users to describe methods, and probably nearly useless for expert users to apply this kind of procedure

if they knew the methods in the required detail, a practical implementation of this procedure would require a pre-collected knowledge base on the methods. Although several attempts made to create such knowledge bases have been cited in this paper, it should be referred that there are so far no consensual descriptions regarding the characteristics of many methods (and there may be even doubt on whether any common set of characteristics may adequately represent all the different methods). However, as the experiment has shown, even without an *a priori* agreed upon core knowledge, it is not impossible for suitable characteristics knowledge (required for the procedure) to be generated in such a way that gives internally coherent (but not necessarily optimal) results.

The sequence of activities to be performed requires elicitation of method and situation characteristics before the MCDM procedure can be applied. It is possible for information about method characteristics to be reused, and after such a body of knowledge is built posterior applications of the procedure could be easier. Notice however that the synergistic relationship between method needs and situation capabilities (and likewise between method abilities and situation requirements) means that may be hard to consider each independently (and there is an obvious connection between familiarity with the method evaluation characteristics domain and the ability to perform correct situation evaluation).

Although two different MCDM approaches were shown, we don't claim either should be the specific way to perform the matching. Many others could be used instead, each with different requirements in terms of information and expertise and providing (possibly) slightly differing results, and opting between them should take into account the preferences of the evaluators, both in terms of required information and of the information provided by the procedure. The study of the MCDM procedure most adequate to this approach, and the operationalization of the approach with a suitable artifact are research avenues being considered.

Finally, a source of the discrepancies between group evaluation of the situation may be not exactly disagreement on the situation characteristics but whether some characteristics should be interpreted as considered as evaluation criteria, or should be considered as constraint criteria. The difference between the two is that on a constraint criteria, to have least a certain value or set of values is required for a method to be considered, whereas on an evaluation criteria/goal criteria each method is allowed to have any values on that criteria but is penalized on the evaluation the farther away it is from the goal. This was particularly noticeable for the "time" criteria, where a group seemed to define it as a constraint (is it possible to use this method in the time given?) rather than as a fit (which time should optimally the evaluation take?). The semantics associated with the criteria should be made explicit, otherwise this kind of disagreement may prevent the results to be adequate for its users. Additional future work will also need to focus on interaction design to support this knowledge elicitation.

References

Amir, Amihood, Eran Chencinski, Costas Iliopoulos, T Kopelowitz, and H Zhang. 2008. "Property Matching and Weighted Matching." *Theoretical Computer Science* 395 (2-3): 298–310. doi:10.1016/j.tcs.2008.01.006. http://linkinghub.elsevier.com/retrieve/pii/S030439750800042X.

Belton, Valerie, and Jacques Pictet. 1997. "A Framework for Group Decision Using a MCDA Model: Sharing, Aggregating or Comparing Individual Information?" *Journal of Decision Systems* 6: 283–303.

Brans, J.P., Philippe Vincke, and Bertrand Mareschal. 1986. "How to Select and How to Rank Projects: The Promethee Method." *European Journal of Operational Research* 24 (2) (February): 228–238. doi:10.1016/0377-2217(86)90044-5. http://linkinghub.elsevier.com/retrieve/pii/0377221786900445.

Card, Stuart, Thomas P. Moran, and Allen Newell. 1986. *The Psychology of Human Computer Interaction*. Lawrence Erlbaum Associates.

Edwards, Ward. 1977. "How to Use Multiattribute Utility Measurement for Social Decisionmaking." *Systems, Man and Cybernetics, IEEE Transactions On* 7 (5): 326 – 340.

Ehrgott, Mathias, and Xavier Gandibleux, eds. 2002. *Multiple Criteria Optimization: State of the Art Annotated Bibliographic Surveys*. Boston: Kluwer Academic Publishers.

Ferre Grau, Xavier, and Nigel Bevan. 2011. "Usability Planner : A Tool to Support the Process of Selecting Usability Methods." In *13th IFIP TCI3 International Conference on Human-Computer Interaction – Interact 2011*, ed. Pedro Campos, Nicholas Graham, Joaquim Jorge, Nuno Nunes, Philippe Palanque, and Marco Winckler, 652–655. Lisbon: Springer Verlag. doi:10.1007/978-3-642-23768-3_105.

Figueira, José Rui, Vincent Mousseau, and Bernard Roy. 2005. "Electre Methods." In *Multiple Criteria Decision Analysis: State of the Art Surveys*, ed. J. Figueira, S. Greco, and M. Ehrgott, 133–162. New York, New York, USA: Springer.

Ghomari, Leila, and Abdessamed Réda Ghomari. 2009. *A Comparative Study: Syntactic Versus Semantic Matching Systems*. *2009 International Conference on Complex, Intelligent and Software Intensive Systems*. IEEE. doi:10.1109/CISIS.2009.75. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=5066864.

Kujala, Sari, Virpi Roto, Kaisa Väänänen-Vainio-Mattila, and Arto Sinnelä. 2011. "Identifying Hedonic Factors in Long-term User Experience." In *Proceedings of the 2011 Conference on Designing Pleasurable Products and Interfaces - DPPI '11*, 1. New York, New York, USA: ACM Press. doi:10.1145/2347504.2347523. http://dl.acm.org/citation.cfm?doid=2347504.2347523.

Lahdelma, Risto, and Pekka Salminen. 2001. "SMAA-2: Stochastic Multicriteria Acceptability Analysis for Group Decision Making." *Operations Research* 49 (3) (May 1): 444–454. doi:10.1287/opre.49.3.444.11220. http://or.journal.informs.org/content/49/3/444.

Law, Chris M., Ji Soo Yi, Young Sang Choi, and Julie A. Jacko. 2007. "A Systematic Examination of Universal Design Resources: Part 1, Heuristic Evaluation." *Universal Access in the Information Society* 7 (1-2) (October 24): 31–54. doi:10.1007/s10209-007-0100-1. http://www.springerlink.com/index/10.1007/s10209-007-0100-1.

Law, Effie Lai-chong, Mark Springett, and Marco Winckler. 2009. *Maturation of Usability Evaluation Methods : Retrospect and Prospect Final Reports of MAUSE*. Toulouse, France: IRIT Press.

Mareschal, Bertrand, and Yves De Smet. 2009. "Visual PROMETHEE: Developments of the PROMETHEE & GAIA Multicriteria Decision Aid Methods." In *2009 IEEE International Conference on Industrial Engineering and*

Engineering Management, 1646–1649. Ieee. doi:10.1109/IEEM.2009.5373124. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5373124.

Medaglia, Andrés L., and Shu-Cherng Fang. 2003. "A Genetic-based Framework for Solving (multi-criteria) Weighted Matching Problems." *European Journal of Operational Research* 149 (1): 77 – 101. doi:10.1016/S0377-2217(02)00484-8. http://www.sciencedirect.com/science/article/pii/S0377221702004848.

Miettinen, Kaisa, and Marko M. Makela. 2002. "On Scalarizing Functions in Multiobjective Optimization." *OR Spectrum* 24 (2) (May 1): 193–213. doi:10.1007/s00291-001-0092-9. http://www.springerlink.com/Index/10.1007/s00291-001-0092-9.

Nielsen, Jakob. 1994. "Usability Inspection Methods." In *Conference Companion on Human Factors in Computing Systems - CHI '94*, 413–414. New York, New York, USA: ACM Press. doi:10.1145/259963.260531. http://portal.acm.org/citation.cfm?doid=259963.260531.

Nielsen, Jakob, and Rolf Molich. 1990. "Heuristic Evaluation of User Interfaces." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Empowering People - CHI '90*, 249–256. New York, New York, USA: ACM Press. doi:10.1145/97243.97281. http://portal.acm.org/citation.cfm?doid=97243.97281.

Nielsen, Jakob, and Darrell Sano. 1995. "SunWeb: User Interface Design for Sun Microsystem's Internal Web." *Computer Networks and ISDN Systems* 28 (1-2) (December): 179–188. doi:10.1016/0169-7552(95)00109-7. http://linkinghub.elsevier.com/retrieve/pii/0169755295001097.

Norman, Geoff. 2010. "Likert Scales, Levels of Measurement and the 'Laws' of Statistics." *Advances in Health Sciences Education : Theory and Practice* 15 (5) (December): 625–32. doi:10.1007/s10459-010-9222-y. http://www.ncbi.nlm.nih.gov/pubmed/20146096.

Silva, Catarina Ferreira, Paulo Rupino Cunha, Parisa Ghodous, and Paulo Melo. 2010. "The Semantic Side of Service-Oriented Architectures." In *Semantic Enterprise Application Integration for Business Processes ServiceOriented Frameworks*, ed. Gregoris Mentzas and Andreas Friesen, 90–104. IGI Global. doi:10.4018/978-1-60566-804-8.ch005. http://www.igi-global.com/bookstore/chapter.aspx?TitleId=37934.

Siskos, Y., E. Grigoroudis, and N. Matsatsinis. 2005. "UTA Methods." In *Multiple Criteria Decision Analysis: State of the Art Surveys*, ed. J. Figueira, S. Greco, and M. Ehrgott, 297–344. New York, New York, USA: Springer.

Stroulia, Eleni, and Yiqiao Wang. 2005. "Structural and Semantic Matching for Assessing Web-Service Similarity." *International Journal of Cooperative Information Systems* 14 (4) (December): 407–437. doi:10.1142/S0218843005001213. http://www.worldscinet.com/ijcis/14/1404/S0218843005001213.html.

Wixon, Dennis, Karen Holtzblatt, and Stephen T. Knox. 1990. "Contextual Design: An Emergent View of System Design." In *Proceedings of the ACM CHI 90 Human Factors in Computing Systems Conference 1990*, ed. Jane Carrasco and John Whiteside. Seattle, Wa.

Woolrych, Alan, Kasper Hornbæk, Erik Frøkjær, and Gilbert Cockton. 2011. "Ingredients and Meals Rather Than Recipes: A Proposal for Research That Does Not Treat Usability Evaluation Methods as Indivisible Wholes." *International Journal of Human-Computer Interaction* 27 (10) (October): 940–970. doi:10.1080/10447318.2011.555314. http://www.tandfonline.com/doi/abs/10.1080/10447318.2011.555314.

Yarbus, Alfred Lukyanovich. 1967. *Eye Movements and Vision*. New York, USA: Plenum Press.