The role of non-transparent matching methods in avoiding preference reversals in the evaluation of health outcomes

Fernando I. Sánchez Martínez
Universidad de Murcia, Spain

Jose Luis Pinto Prades*
Glasgow Caledonian University, UK

José María Abellán Perpiñán
Universidad de Murcia, Spain

Jorge E. Martínez Pérez
Universidad de Murcia, Spain

*** Corresponding author**

Yunus Centre for Social Business & Health
Institutes for Applied Health Research and Society & Social Justice Research
3rd Floor, Buchanan House
Glasgow Caledonian University
Cowcaddens Road
Glasgow, G4 0BA

1. INTRODUCTION

Two major classes of preference elicitation methods are Matching and Choice. In Matching Methods, respondents are asked to establish indifference between two options. For example, Sumner and Nease (2001) asked subjects to specify the missing value in (30 years, ___ migraine days per month) so that they were indifferent between this option and (20 years, 4 migraine days per month). In Choice subjects are given (usually) two options where all parameters are fixed and subjects have to state their preferences between the two. Under the assumption that there is a unique utility function (utility invariance) that subjects apply to Matching and Choice, we would expect that both tasks would reveal the same preferences. However, very often this is not the case. Sumner and Nease (2001) also asked subjects to choose between (30 years, 10 migraine days per month) and (20 years, 4 migraine days per month). In Choice, most people preferred (30 years, 10 migraine days per month) while in Matching most people stated a number of migraine days per month lower than 10, that is, they (implicitly) preferred (20 years, 4 migraine days per month). This is an example of a Preference Reversal. It suggests that Choice and Matching seem to reveal opposite preferences and it is not clear which are the "true" preferences. This is an important topic since a systematic discrepancy between Matching and Choice in the evaluation of health states violates procedure invariance, one of the cornerstones of rational choice. The objective of this paper is to suggest how they can be avoided or, at least, mitigated.

We study if there are elicitation methods that reduce or eliminate the discrepancy between Matching and Choice in the domain of the evaluation of health outcomes. In health economics this topic is especially relevant since, in recent years, there has been a renewed interest in the use of preference elicitation methods based on Choice (e.g. Discrete Choice Experiments –DCE) to elicit utilities for health states (Clark et al., 2014). There is some evidence suggesting that utilities elicited with DCE and utilities elicited with methods based on Matching (like Standard Gamble or Time Trade-Off) are different (Bansback et al. 2012). The literature suggests that some matching methods can produce results more consistent with choices. However, the evidence is relatively scarce and it is totally absent in the domain of health outcomes.

We present the results of an experiment that compares choices and several matching procedures. All matching methods are choice-based, that is, they match two options using converging sequences of choices (this Choice-Based Matching (CBM) approach will be explained later). Our main finding is that Preference Reversals are avoided when CBM methods hide respondents the final goal of the sequence (what we call "non-transparent" or "opaque" methods). That is, when subjects do not see that each choice is part of a sequence aimed at establishing indifference between options. This seems to confirm Fischer et al's (1999) task-goal hypothesis that has not been tested previously in health economics. We suggest that methods like the Time Trade-Off or the Standard Gamble can be improved using non-transparent sequences of choices. In the next section we review the explanations provided in the literature about the Choice-Matching discrepancy. Sections 3 and 4 contain the methodology and the results, respectively, of our study. The paper ends with some conclusions.

2. THE CHOICE-MATCHING DISCREPANCY

2.1. The phenomenon

The Choice-Matching Discrepancy was first observed using monetary gambles (Lichtenstein and Slovic, 1971). Subjects were asked to choose between two lotteries, both with two states of the world: success and failure. Failure implies a 0 gain and success a positive gain. Both lotteries have similar expected values, but in one of them, subjects have a high chance of a small gain; and in the other one, there is a low chance of a large gain. The first lottery is usually called the "P-bet" and the second the "$-bet". Subjects are asked to state their preference between the two lotteries with two different tasks, namely, monetary equivalence (a matching task) and direct choice. In the monetary equivalence task, subjects are asked to state the certain monetary amount that has the same value as the lottery. In the choice task subjects have to choose between the P-bet and the $-bet. The Choice-Matching Discrepancy is normally characterized by people giving a higher monetary equivalent to the $-bet than to the P-bet but choosing the P-bet over the $-bet in the straight choice.

Butler and Loomes (2007) also found a discrepancy between choice and matching in another type of matching task. In this case, the common currency they used to value two lotteries (one P-bet and one $-bet) was what they called a "Reference lottery" (R-bet). This lottery is characterized by having a 0 outcome if unsuccessful but the payoff if successful is higher than the highest payoff in the $-bet. The matching task is called "Probability Equivalent" since subjects had to state the probabilities in the R-bet such that they are indifferent between the R-bet and the $-bet and the same for the P-bet. They found the opposite asymmetry, that is, P-bet$\succ$$-bet in matching and $-bet$\succ$P-bet in choice.

2.2. Explanation of the phenomenon

This phenomenon has been explained in terms of a psychological principle called *Compatibility between stimulus and response*, originally observed by Fitts and Seeger (1953) in sensory tasks. According to this principle the respondent weights more heavily the characteristics of the stimulus that are more compatible with the response. Several effects have been explained using this general principle. One is *scale compatibility*. According to this principle, it is the similarity between stimulus and response scales that leads subjects to overweight the compatible attribute. For example, in the Sumner and Nease (2001) study the response scale was days of migraine. Scale Compatibility implies that in the Matching exercise subjects give too much weight to "migraine days per month" since this is the response scale. In the case of the P-bet and the $-bet, the response scale is money in the matching task so the monetary outcomes of the lotteries are overweighted. This favours the $-bet lottery, since it has the highest monetary price. The pattern in Butler and Loomes (2007) could also be explained by *scale compatibility*, that is, overweighting of the attribute that is used to reach indifference in the valuation task, probabilities in their case.

A second form of compatibility that has been suggested to contribute to the Choice-Matching Discrepancy is called *strategy compatibility*. Strategy compatibility implies that the strategies that subjects follow in each task are different. In choices subjects may follow qualitative strategies because choice is a qualitative strategy; for example, a choice may be decided using lexicographic principles or aspiration levels. The implication is that in choice subjects focus mainly on the most important (or prominent) attribute. This is called the *Prominence Hypothesis*, namely, in choices the most important attribute receives higher weight than in matching. Empirical results suggest that the probability of winning is the prominent attribute in the monetary lotteries commonly used in experiments (Slovic et al. 1990). In summary, Scale Compatibility leads to an overvaluation of the $-bet in Matching while Strategy Compatibility enhances the attractiveness of the P-bet. This leads to the well-known result that P-bet$\succ$$-bet in choice and $-bet$\succ$P-bet in matching.

Fischer et al (1999) provide another explanation of the choice-matching discrepancy. They propose the *task-goal hypothesis*. According to this hypothesis the prominent attribute is weighted more heavily in tasks whose perceived goal is to differentiate between alternatives than in tasks whose goal is to equate alternatives. The reason is that to differentiate only requires to rank-order the alternatives, which is naturally compatible with choosing the alternative that is superior in the prominent attribute. To equate requires making trade-offs between attributes, which is naturally compatible with giving some weight to all the attributes. The implication is that the prominent attribute will receive more weight in response tasks whose perceived goal is to differentiate between alternatives (Choice) than in tasks whose perceived goal is to equate between alternatives (Matching).

2.3. Classifying Choice-Based Matching methods.

In this paper we will use several CBM methods that, depending on the way Preference Reversals are explained, are or are not expected to avoid the Preference Reversal phenomenon. We proceed to describe the characteristics of those methods and we next explain when they are expected to avoid Preference Reversals.

2.3.1. Standard Matching vs. Choice-Based Matching.

Assume we have two objects (A, B) with two attributes [X, Y], that is, object A is characterized by (X_A, Y_A) and object B by (X_B, Y_B). We want to estimate a combination of attributes such that A and B are equally attractive. One possibility is to fix three of the four attributes and the subject has to specify the value of the omitted attribute that makes him/her indifferent between A and B. For example, the subject has to state the missing value (?) in an open question so that (X_A, Y_A) and (X_B, ?) are equally preferred. Assume the value is Y_B^* then (X_A, Y_A)$\sim$(X_B, Y_B^*). We call this Standard Matching and one example is the already mentioned Sumner and Nease (2001) paper.
Choice-Based Matching does not reach indifference with an open question but with a sequence of choices. This process generates an interval where the indifference point is located. Assume that (X_A, Y_A) $\succ$ (X_B, Y_B^1) but (X_A, Y_A) $\prec$ (X_B, Y_B^2). We then know that the

value of $\mathrm{Y_B}$ that will make the two options equally attractive will be in the interval $[Y_B^1, Y_B^2]$ Researchers may take the middle point of the interval as the indifference (matching) point or they may ask an open question with the matching point constrained by the interval where indifference was located.

2.3.2. Iterative vs. non-iterative CBM.

There are different variants within CBM methods. They can be iterative or non-iterative. In an iterative method, the choice that the subject is presented depends on his/her response to a previous choice. For example, assume that the subject says that $(\mathrm{X_A}, \mathrm{Y_A}) \prec (\mathrm{X_B}, Y_B^1)$. In an iterative method, the subject would be then offered a choice between $(\mathrm{X_B}, Y_B^2)$ and $(\mathrm{X_A}, \mathrm{Y_A})$ for $Y_B^2 < Y_B^1$, but he/she would never be presented a choice between $(\mathrm{X_B}, Y_B^2)$ and $(\mathrm{X_A}, \mathrm{Y_A})$ for $Y_B^2 > Y_B^1$. In non-iterative methods, questions are set up in advance, and subjects respond to all of them independently of their responses to previous choices. Even if the subject says that $(\mathrm{X_A}, \mathrm{Y_A}) \prec (\mathrm{X_B}, Y_B^1)$ the choice between $(\mathrm{X_B}, Y_B^2)$ and $(\mathrm{X_A}, \mathrm{Y_A})$ could be offered for $Y_B^2 > Y_B^1$ if this choice was included in the set of choices that was established before the subject was interviewed. In health economics, CBM methods (Time Trade-Off and Standard Gamble) are usually iterative. Non-iterative methods are not very common in health economics but the Multiple Price List method so widely used to elicit risk preferences (Holt and Laury, 2002) is non-iterative. One problem with non-iterative methods is that subjects can give inconsistent responses. However, for those who do not make mistakes, an indifferent point (or interval) can be estimated. Finally, depending on the rules to generate the stimuli from one choice to another (e.g, how we choose), we can further split iterative methods in Titration, Bisection, “Ping-pong” and so on.

2.3.3. Transparent vs. non-transparent CBM

When subjects can easily observe that there is a link between the choices of a converging sequence, we say that the method is “transparent”. If it is difficult for people to observe this link, we talk about a “non-transparent” method. Let us see an example. Assume we want to estimate the utility of *N* health states. This defines *N* converging sequences of choices in CBM. The usual way of eliciting preferences would be to start with a certain health state (say #1), apply the corresponding converging sequence until indifference is reached and then move to a different health state and start again with another converging sequence of choices. All choices for health state #1 have 3 attributes that are constant and one is moving up and down. Some subjects may quickly realize of the kind of “game” they are playing. However, an alternative way of eliciting preferences would be to ask subjects to make one choice from converging sequence #1, then one choice from sequences #2, #3, …, #*N*, before returning to sequence #1. In this way, when the subject is presented with the second choice of converging sequence #1 we hope she cannot see that this choice is related to a choice that she made *N* choices ago. This is the method used by Fischer et al. (1999) that they coined as *Hidden Choice-Based Matching* (HCBM). This would be a non-transparent method. While the iterative or non-iterative nature of a CBM method is objective, it is not the same in the case of transparent vs. non-transparent methods. For example, in the HCBM method, we can expect the method being less non-transparent the less converging sequences we use. However, we would expect that the majority of people may find more difficult to observe that the choice belong to a sequence with HCBM than with the traditional method of using sequences of choices for the same health state until indifference is reached.

Once we have presented the different ways of applying Matching methods to elicit preferences, can we predict which of them will eliminate the Choice-Matching discrepancy? We explain this next.

2.4. Avoiding the discrepancy

We would expect Standard Matching being very influenced by Scale Compatibility since the subjects are asked to respond using a certain scale. This will lead to overweighting of the response scale. Also, the task in Standard Matching is quantitative and in Choice is qualitative. This is another reason why Matching and Choice will diverge, since Choice will overweight the more prominent attribute and the P-bet will be valued higher than the $-bet. An interesting case is when Standard Matching is done using the prominent attribute, that is, probability in monetary lotteries. Slovic et al (1990) find that in that case preferences elicited using Standard Matching and Choice coincide. They interpret that as the consequence of two biases (Compatibility and Prominence) with the same strength. They do not interpret that as evidence of consistency derived from well-defined preferences. They support that interpretation in another finding, that is, they also applied Standard Matching using the non-prominent attribute (money) and in that case subjects showed much weaker preferences for the P-bet. However, Fischer and Hawkins (1993) evidence is that Prominence is stronger than Scale Compatibility. In any case, a Standard Matching task has two features (scale used, quantitative task) that makes it very different from Choices and prone to producing Preference Reversals.

If Scale Compatibility or Strategy Compatibility are the reasons behind the Choice-Matching disparity, it would be enough to move from standard matching to CBM to avoid Preference Reversals. If each of the tasks in CBM is perceived as an independent choice, preferences elicited with CBM and direct choices should converge. The task is now qualitative (choose A or B) and the question does not involve the use of a scale. However, if the Task-Goal is the correct explanation of the Choice-Matching disparity, it will not be enough to use CBM methods to avoid Preference Reversals. The key issue to avoid Preference Reversals is that subjects should not perceive that the objective of the CBM procedure is equating between options. Only non-transparent methods will avoid Preference Reversals.

In this paper we will use CBM methods that are iterative+transparent, iterative+non-transparent, non-iterative+transparent and non-iterative+non-transparent. If the combination of Scale and Strategy Compatibility explains the Choice-Matching discrepancy, all methods should eliminate the discrepancy since they are all choice-based. If Task-Goal is the right explanation, non-transparent methods will reduce the discrepancy more than non-transparent methods. If none of the methods reduce the Choice-Matching disparity new explanations would be needed.

3. METHODS

3.1. Participants

We recruited a total of 250 undergraduate students at the University of Murcia (Spain). Participants were randomly allocated[1] to one of five groups, which differed in the type of elicitation mode used in the CBM procedure. The sessions took place in the Lab of the Faculty of Economics and Business of the University of Murcia, under the supervision of members of the research team. A total of 14 sessions were held with less than 25 students in each session. Students were paid €15 for their participation. Sessions took about 40 minutes to complete.

3.2. Gambles and tasks

The sessions involved two types of tasks: straight choices between two gambles, a P-bet and a $-bet, and valuations of each gamble by means of a CBM procedure. Each task was repeated three times.

Two pairs of lotteries were used (see Table 1). In each lottery one outcome was a chronic health condition described in terms of an EQ-5D[2] health state, and the other outcome was immediate death. The so-called P-bet lotteries (A and C) offered the individuals a large probability of being in a bad health state for the rest of their lives, whereas the $-bet (lotteries B and D) gave them a low probability of living in a better health state but a higher risk of death. The two sets of paired lotteries (A-B, C-D) had similar expected utilities according to the EQ-5D-3L Spanish TTO tariff (Badia et al. 2001).

Table 1. Lotteries used in the study

	Pair 1	EU*	Pair 2	EU*
P-bet	**A**: (12231, 0.95; *Death*)	0.21	**C**: (22223, 0.8; *Death*)	0.12
$-bet	**B**: (11221, 0.3; *Death*)	0.24	**D**: (12221, 0.2; *Death*)	0.14

[1] In order to do that, we included all subjects who volunteered for the experiment in a data base and they were allocated to one of the five CBM methods using a random number generator. Since we wanted to have exactly the same number of subjects in each version, the number corresponding to each CBM method was omitted once we had 50 subjects allocated to that procedure. When they introduced their National Identity Card number in the computer they were allocated to only one of the methods.

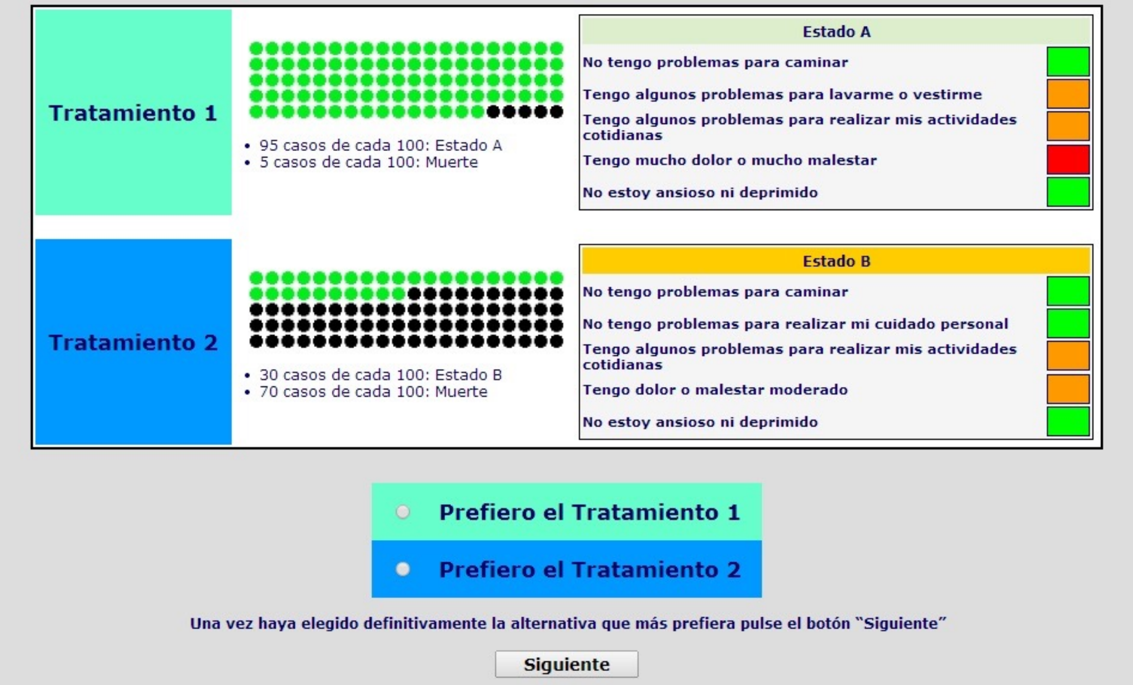
[2] The EQ-5D descriptive system includes five dimensions: mobility, self-care, usual activities, pain/discomfort and anxiety/depression. Each dimension has three levels: no problems, some problems, severe problems. State 12231, for instance, describes the condition of an individual who has no problems in walking about nor is anxious or depressed, but who has some problems washing or dressing herself and performing usual activities, and who has extreme pain or discomfort too.

* Expected Utility according to Spanish tariff: U(12231)=0.219; U(11221)=0.816; U(22223)=0.141; U(12221)=0.682.

The scenarios described a hypothetical situation where subjects had to choose between two treatments. Otherwise, they were told that they would be die in a few days. Given that the default option or reference point was certain death (utility=0), each treatment was in the gain domain. So a P-bet is a treatment that offers a large probability of a small gain (e.g. from 0 to 0.219 or 0.141) while the $-bet offers a smaller chance of a bigger gain (e.g. from 0 to 0.816 or 0.682).

Visual aids were used to represent the probabilities of success and failure in each treatment. An example can be seen in Figure 1 that represents a direct choice between lotteries A and B.

Figure 1. Example of a direct choice task (Lottery A vs. Lottery B).



Tratamiento 1

- 95 casos de cada 100: Estado A
- 5 casos de cada 100: Muerte

Estado A	
No tengo problemas para caminar	🟩
Tengo algunos problemas para lavarme o vestirme	🟧
Tengo algunos problemas para realizar mis actividades cotidianas	🟧
Tengo mucho dolor o mucho malestar	🟥
No estoy ansioso ni deprimido	🟩

Tratamiento 2

- 30 casos de cada 100: Estado B
- 70 casos de cada 100: Muerte

Estado B	
No tengo problemas para caminar	🟩
No tengo problemas para realizar mi cuidado personal	🟩
Tengo algunos problemas para realizar mis actividades cotidianas	🟧
Tengo dolor o malestar moderado	🟧
No estoy ansioso ni deprimido	🟩

○ Prefiero el Tratamiento 1

○ Prefiero el Tratamiento 2

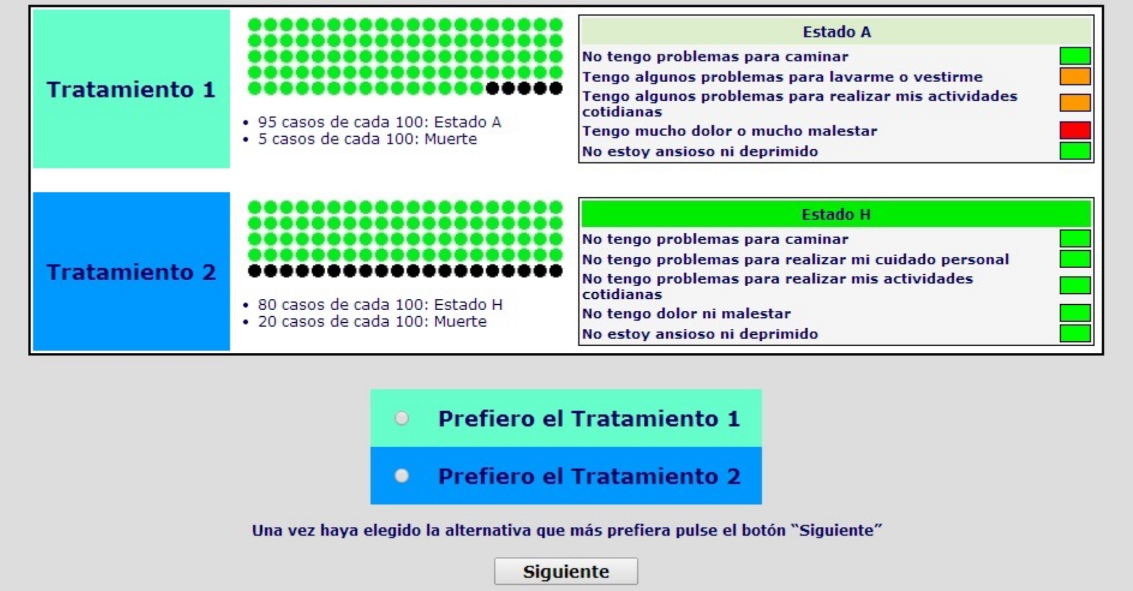
Una vez haya elegido definitivamente la alternativa que más prefiera pulse el botón "Siguiente"

The valuation task consisted of a sequence of choices between each of the lotteries and a reference gamble (R), whose best outcome was full health and the worst was death:

R: (*Full health*, *p*; *Death*)

That is, we used a variant of the lottery equivalent procedures (McCord and de Neufville, 1986) called *Probability Lottery Equivalent* (PLE), since the equivalence between the gambles is obtained changing the probability in one of the lotteries (the R gamble, in our case), as in Butler and Loomes (2007) study. This technique has been previously used to estimate the Spanish SF-6D scoring algorithm (Abellan et al., 2012). An example of a PLE question can be seen in Figure 2.

Figure 2. Example of a PLE question.



Tratamiento 1

- 95 casos de cada 100: Estado A
- 5 casos de cada 100: Muerte

Estado A	
No tengo problemas para caminar	🟩
Tengo algunos problemas para lavarme o vestirme	🟧
Tengo algunos problemas para realizar mis actividades cotidianas	🟧
Tengo mucho dolor o mucho malestar	🟥
No estoy ansioso ni deprimido	🟩

Tratamiento 2

- 80 casos de cada 100: Estado H
- 20 casos de cada 100: Muerte

Estado H	
No tengo problemas para caminar	🟩
No tengo problemas para realizar mi cuidado personal	🟩
No tengo problemas para realizar mis actividades cotidianas	🟩
No tengo dolor ni malestar	🟩
No estoy ansioso ni deprimido	🟩

○ Prefiero el Tratamiento 1

○ Prefiero el Tratamiento 2

Una vez haya elegido la alternativa que más prefiera pulse el botón "Siguiente"

Indifference was achieved by changing the probability *p* in lottery R. The possible values for probability *p* were predetermined before the matching procedure started. It was necessary to define in advance the values of *p* given that in non-iterative methods the subject is asked a predetermined number of questions independently of her responses to previous questions. In order to make matching tasks between the five methods as similar as possible, we decided to adopt the same predetermined values in iterative and non-iterative methods. In each of the four lotteries nine different values of *p* were used (see Table 2). In iterative methods subjects were only asked a subset of those nine values while in non-iterative they were asked the nine values. Indifference was not allowed neither in straight choices (A vs. B; C vs. D) nor in PLE.

Table 2. Reference gambles in Probability Equivalent questions.

	A: (12231, 0.95; D)	B: (11221, 0.3; D)	C: (22223, 0.8; D)	D: (12221, 0.2; D)
R:	(11111, 0.1; D) (11111, 0.2; D) (11111, 0.3; D) (11111, 0.4; D) (11111, 0.5; D) (11111, 0.6; D) (11111, 0.7; D) (11111, 0.8; D) (11111, 0.9; D)	(11111, 0.03; D) (11111, 0.06; D) (11111, 0.09; D) (11111, 0.12; D) (11111, 0.15; D) (11111, 0.18; D) (11111, 0.21; D) (11111, 0.24; D) (11111, 0.27; D)	(11111, 0.08; D) (11111, 0.16; D) (11111, 0.24; D) (11111, 0.32; D) (11111, 0.40; D) (11111, 0.48; D) (11111, 0.56; D) (11111, 0.64; D) (11111, 0.72; D)	(11111, 0.02; D) (11111, 0.04; D) (11111, 0.06; D) (11111, 0.08; D) (11111, 0.10; D) (11111, 0.12; D) (11111, 0.14; D) (11111, 0.16; D) (11111, 0.18; D)

3.3. Choice-based matching methods

Five different types of matching methods were used: two of them are iterative and transparent; one is iterative but opaque; another one is non-iterative and transparent and the last one is non-iterative and non-transparent. We describe each of them in turn.

The transparent iterative methods we used are *Bisection* and a modified version of the *Ping-pong* procedure. In both methods the first value of the matching parameter was randomly chosen amongst the nine potential values of *p*. Assume it was p_3. This generated two potential intervals where the indifference point had to be located, namely, [0-p_3] and [p_3-p^i_{max}]. For example, in the comparison between lottery (12231, 0.95; *Death*) and (11111, p_3;Death), we have that p_3=0.3. If the respondent chose lottery (11111, 0.3; Death), we knew that the second stimulus (value of *p*) had to be 0.1 or 0.2. If the respondent chose lottery (12231, 0.95; *Death*), the second stimulus had to be one value of the set {0.4, 0.5, 0.6, 0.7, 0.8, 0.9}. The difference between both methods (Bisection and Ping-pong) was how they select the value of *p* in the second question. Assume the subject prefers lottery (12231, 0.95; *Death*) to lottery (11111, 0.3; Death). In the bisection method, the second value of *p* would be the value closest to the middle point of the interval [p_3-p^i_{max}], which in this case it would be p_7. The second choice would be then (12231, 0.95; Death) vs (11111, 0.7; Death). In the ping-pong method, the second value of *p* would be located at the other end of the interval opposite to p_3. In this example, it would be p_9. The second choice would be (12231, 0.95; Death) vs (11111, 0.9; Death). In both cases, the process goes on until the indifference point is located within one of the ten intervals $[\overline{p_L^i - p_U^i}]$ defined. At this point, the process stops. The lower limit of the interval (p_L^i) is the highest value of *p* in lottery R(*Full health*, *p*; *Death*) for which the individual prefers the lottery *i* to *R*; and the upper end of the interval (p_U^i) is the lowest value of *p* in lottery R for which the subject prefers the lottery *R* to *i*. For example, if (12131, 0.95; Death)$\succ$(11111, 0.6: Death) but (12231, 0.95; Death)$\prec$(11111, 0.7: Death) then p_L^i=0.6 and p_U^i=0.7 so $[\overline{p_L^i - p_U^i}]$ = $[\overline{0.6 - 0.7}]$.

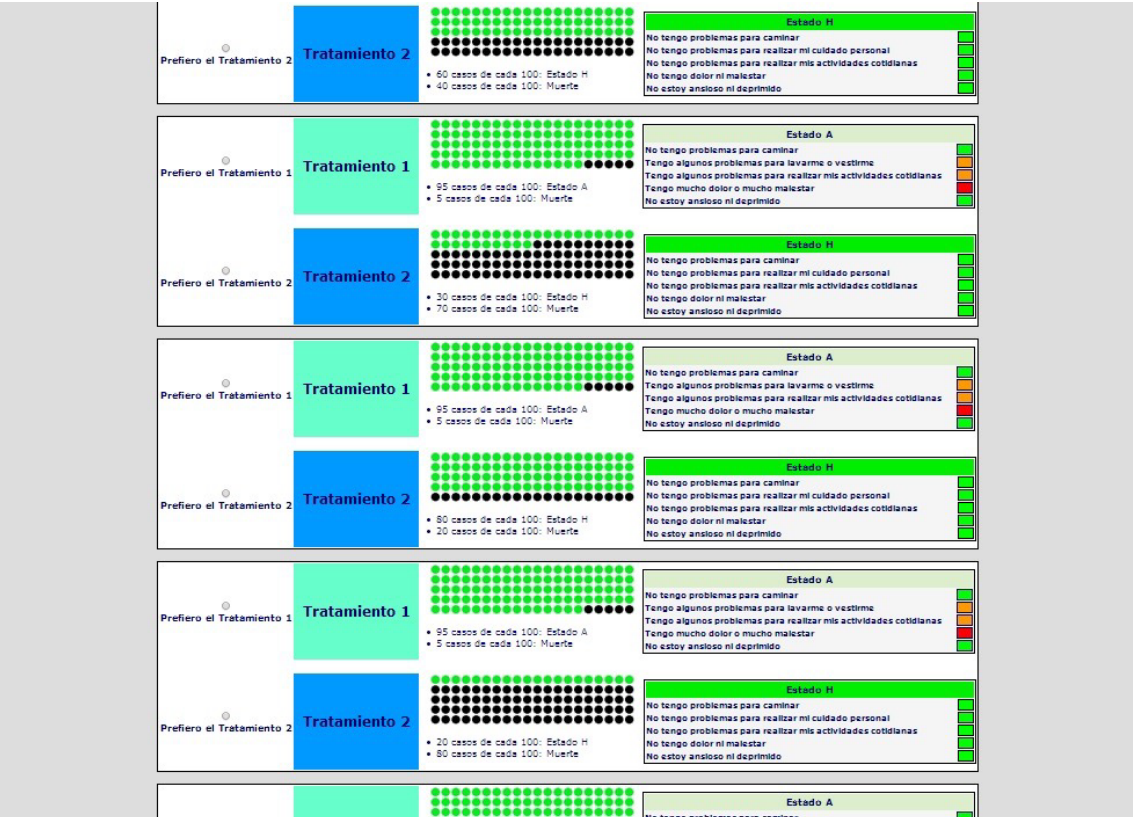
The non-transparent or 'opaque' iterative procedure applied was the HCBM proposed by Fischer et al. (1999). The HCBM was applied using the bisection method but separating the choices regarding each particular lottery by the iterations of other lotteries. Thus, subjects make one choice from iteration process belonging to lottery A, then one choice for lottery B, one for lottery C and one for lottery D before returning to the sequence of lottery A.[3] For example, a hypothetical sequence could have been as follows: first choice between (12231, 0.95; *Death*) and (11111, 0.3; Death), the second choice between (11221, 0.3; *Death*) and (11111, 0.09; Death), the third choice between (22223, 0.8; *Death*) and (11111, 0.56; Death) and the fourth choice between (12221, 0.2; *Death*) and (11111, 0.16; Death)[4]. Assuming that in all cases the reference lottery was preferred, the fifth to eighth choices were between (12231, 0.95; *Death*) vs (11111, 0.2; Death), (11221, 0.3; *Death*) vs (11111, 0.06; Death), (22223, 0.8; *Death*) vs (11111, 0.32; Death) and (12221, 0.2; *Death*) vs (11111, 0.08; Death).

[3] As Fischer et al. (1999) did in their study, when a sequence converged faster than others, filler choices were added at the end of that sequence, to avoid the possibility that there were only one or two sequences that had not converged at the final stages, thus making the iterative process transparent to the subjects.

[4] In the first sequence of choices, probabilities of the reference lottery (0.3, 0.09, 0.56 and 0.16, respectively) were randomly set amongst the nine potential predetermined values in each case.

The other two methods were non-iterative, that is, subjects had to respond to all possible predetermined choices (i.e. for all values from p_1^i to p_9^i in Table 2). One of these two non-iterative methods was transparent and the other was non-transparent. The first method is a list/table containing all the possible choices for each of the matching sequences (i.e. the four lotteries), which are displayed in random order as it can be seen in Figure 3. We will refer to this procedure as *List*. The second one is a method that resembles (to some extent) a DCE experiment keeping the "essence" of matching, that is, we can obtain the indifference point at the individual level. This procedure, which we will call *Random Binary Choice* (RBC) method, is non-iterative, since it presents the nine potential choices to each subject for each of the four lotteries. Thirty-six choices were randomly presented to each subject. As we have explained, in non-iterative methods the indifference interval $[p_L^i - p_U^i]$ might not be determined if participants make mistakes. This is a problem in *List* and RBC but not in the rest of methods.

Figure 3. Detail of a set of choices in a matching sequence in Group 4 (*List* procedure)



○ Prefiero el Tratamiento 2
Tratamiento 2
- 60 casos de cada 100: Estado H
- 40 casos de cada 100: Muerte

Estado H
No tengo problemas para caminar 🟩
No tengo problemas para realizar mi cuidado personal 🟩
No tengo problemas para realizar mis actividades cotidianas 🟩
No tengo dolor ni malestar 🟩
No estoy ansioso ni deprimido 🟩

○ Prefiero el Tratamiento 1
Tratamiento 1
- 95 casos de cada 100: Estado A
- 5 casos de cada 100: Muerte

Estado A
No tengo problemas para caminar 🟩
Tengo algunos problemas para lavarme o vestirme 🟧
Tengo algunos problemas para realizar mis actividades cotidianas 🟧
Tengo mucho dolor o mucho malestar 🟥
No estoy ansioso ni deprimido 🟩

○ Prefiero el Tratamiento 2
Tratamiento 2
- 30 casos de cada 100: Estado H
- 70 casos de cada 100: Muerte

Estado H
No tengo problemas para caminar 🟩
No tengo problemas para realizar mi cuidado personal 🟩
No tengo problemas para realizar mis actividades cotidianas 🟩
No tengo dolor ni malestar 🟩
No estoy ansioso ni deprimido 🟩

○ Prefiero el Tratamiento 1
Tratamiento 1
- 95 casos de cada 100: Estado A
- 5 casos de cada 100: Muerte

Estado A
No tengo problemas para caminar 🟩
Tengo algunos problemas para lavarme o vestirme 🟧
Tengo algunos problemas para realizar mis actividades cotidianas 🟧
Tengo mucho dolor o mucho malestar 🟥
No estoy ansioso ni deprimido 🟩

○ Prefiero el Tratamiento 2
Tratamiento 2
- 80 casos de cada 100: Estado H
- 20 casos de cada 100: Muerte

Estado H
No tengo problemas para caminar 🟩
No tengo problemas para realizar mi cuidado personal 🟩
No tengo problemas para realizar mis actividades cotidianas 🟩
No tengo dolor ni malestar 🟩
No estoy ansioso ni deprimido 🟩

○ Prefiero el Tratamiento 1
Tratamiento 1
- 95 casos de cada 100: Estado A
- 5 casos de cada 100: Muerte

Estado A
No tengo problemas para caminar 🟩
Tengo algunos problemas para lavarme o vestirme 🟧
Tengo algunos problemas para realizar mis actividades cotidianas 🟧
Tengo mucho dolor o mucho malestar 🟥
No estoy ansioso ni deprimido 🟩

○ Prefiero el Tratamiento 2
Tratamiento 2
- 20 casos de cada 100: Estado H
- 80 casos de cada 100: Muerte

Estado H
No tengo problemas para caminar 🟩
No tengo problemas para realizar mi cuidado personal 🟩
No tengo problemas para realizar mis actividades cotidianas 🟩
No tengo dolor ni malestar 🟩
No estoy ansioso ni deprimido 🟩

Estado A

3.4. Structure of the sessions

The sessions began with an introduction about the experiment. The EQ-5D descriptive system was also briefly explained to the participants and the four health states involved in the lotteries were shown. Subjects were then asked to rate the four states plus the *Death* state on a visual analogue scale, to familiarize participants with the health conditions they had to evaluate.

Subjects were asked to do two types of tasks: choices between paired lotteries (one P-bet and one $-bet) and separate valuations of each lottery by means of the PLE technique (using sequences of choices). In all groups the first task was to choose between lotteries A and B. Then, subjects in groups 1 (*bisection*), 2 (*ping-pong*), and 4 (*List*) were asked to do the valuation (choice-based) task of A and B, followed by the choice between C and D and the valuation of C and D. In groups 3 and 5 (corresponding, respectively, to HCBM and RBC), subjects started with the two straight choices (A vs. B and C vs. D) and then continued with the choice-based valuations of the four lotteries in the manner that has been explained. The same scheme was repeated three consecutive times during the session. For groups 1, 2 and 4, the order in which A and B were valued through the PLE method was randomly determined, and so it was the order between C and D valuations. For groups 3 and 5, the order of appearance of the four lotteries in PLE valuations was set at random.

2.5. Analysis

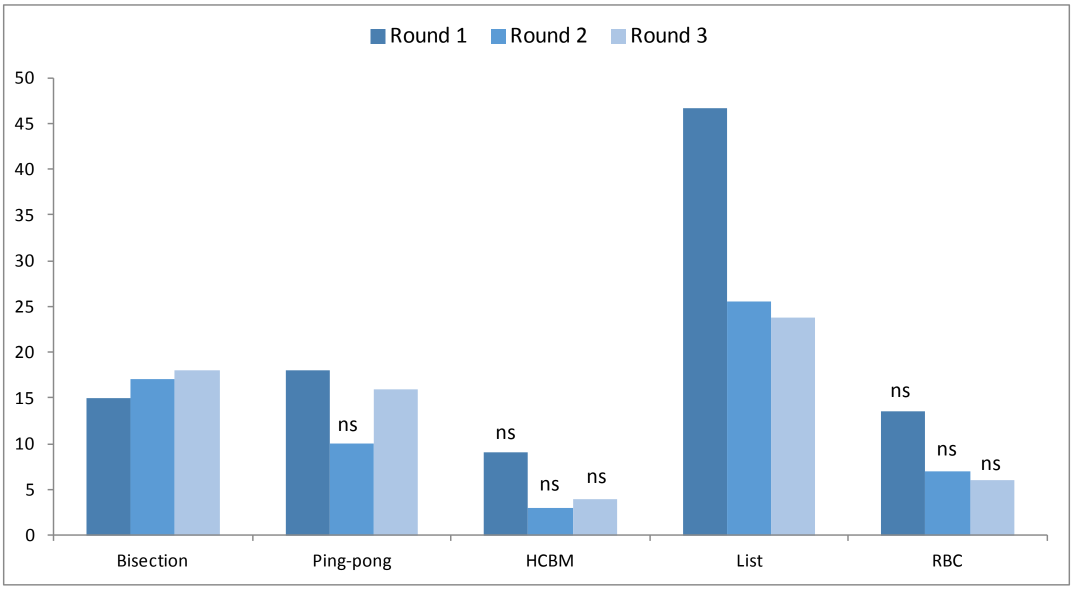
Within each pair of lotteries *i* and *j*, a choice-matching discrepancy exists when *i*≻*j* in a straight choice but PLE valuations imply that *i*≺*j*. Since we did not derive an exact indifference value for the lotteries but only an indifference interval $[\overline{p_L^i - p_U^i}]$ we need to define the procedure to estimate the choice implicit in a PLE task. One option is to assume that the indifference value is in the middle point of the indifference interval. Another option would be to compare the indifference intervals and assume that the PLE valuation task reveals a preference for *i* or *j* only when the two intervals do not overlap. That is, we will say, for instance, that a PLE task implies that *i*≻*j* if $p_L^i > p_U^j$. If p_L^i and p_U^joverlap, no direction of preference can be established. We will present the results using the first method in the main text; the results with the second method are presented in an appendix. The pattern is very similar in both cases as we will show.

4. RESULTS

We present the results for each round separately since preferences seem to change from the first to the third round (see Figure 4). The total number of Preference Reversals was 148 in the first round (Table 3), 102 in the second round (Table 4) and 92 in the third one (Table 5). However, in spite of the reduction in the number of Preference Reversals, they are still highly asymmetric for Bisection, Ping-pong and List in the same direction, namely, the P-bet being more highly preferred in the Valuation than in Choice. Even in the third round (see Table 5), where the lower number of Preference Reversals could be interpreted as better formed preferences, the ratios were 18:0 (Bisection), 17:1 (Ping-pong) and 19:0 (List). That is, an impressive 54:1 in favor of the P-bet in Valuation vs. Choice. However, HCMB and RCB ratios were only 9:5 and 6:1, respectively (p=0.42 and p=0.13, McNemar exact binomial test, 2-sided) for a total of 15:6. This asymmetry in Preference Reversals translates into the P-bet strongly preferred over the $-bet in Bisection, Ping-pong and List. Again, if we only focus on the third round, 51.8% of subjects in transparent methods preferred the P-bet in Choice and 70.7% in Matching, while for HCMB and RBC it was 48.6% vs. 53.6% for Choice vs. Matching, respectively. In summary, for Bisection, Ping-pong and List the results are in agreement with the evidence in Butler and Loomes (2007), that is, the P-bet more highly preferred in Matching.

The common element of the two methods where Preference Reversals almost disappeared (HCBM and RBC) is that they are non-transparent; one is iterative and the other is not. It seems that when subjects are not aware that each choice is part of a sequence, they take each decision using the same principles that they use in straight choices. The method that leads to more discrepancies between valuation and choice is List. It seems that this presentation makes the matching attribute even more salient in the valuation task and probabilities are even more overweighted.

Figure 4. Percentage of responses favoring P-bet in matching minus percentage of responses favoring P-bet in choice.



ns: Statistically non-significant differences, according to Fisher's exact test. In all other cases differences were statistically different from zero at 5% level.

Table 3. Direct choices vs choices implied by the valuation task. Round 1.

		Valuation				A>B//C>D (%)		
	Choice	A>B	A<B	C>D	C<D	Choice	Valuation	p-value(3)
Bisection	A>B//C>D	24	4	15	1	44,0	59,0	0,0051
	A<B//C<D	6	16	14	20			
Ping-pong	A>B//C>D	27	3	18	1	49,0	67,0	0,0009
	A<B//C<D	8	12	14	17			
HCBM	A>B//C>D	25	4	11	2	42,0	51,0	0,0809
	A<B//C<D	6	15	9	28			
List(1)	A>B//C>D	17	0	10	0	30,7	77,3	<0,0001
	A<B//C<D	16	11	25	9			
RBC(2)	A>B//C>D	12	0	7	2	35,6	49,2	0,0433
	A<B//C<D	6	11	4	17			

(1) In group 5 (List), due to inconsistencies, it was not possible to obtain the indifference interval in 6 occasions for pairs A-B and C-D. (2) In group 6 (RBC) the indifference interval could not be identified in 21 occasions for pair A,B and in 20 cases for pair C-D. (3) McNemar's test 2-sided.

Table 4. Direct choices vs choices implied by the valuation task. Round 2.

		Valuation				A≻B//C≻D (%)		
	Choice	A≻B	A≺B	C≻D	C≺D	Choice	Valuation	p-value(3)
Bisection	A≻B//C≻D	28	2	19	0	49,0	66,0	0,0005
	A≺B//C≺D	6	14	13	18			
Ping-pong	A≻B//C≻D	35	1	22	0	58,0	68,0	0,0094
	A≺B//C≺D	0	14	11	17			
HCBM	A≻B//C≻D	28	2	14	2	46,0	49,0	0,5465
	A≺B//C≺D	3	17	4	30			
List(1)	A≻B//C≻D	26	0	17	0	50,0	75,6	<0,0001
	A≺B//C≺D	7	8	15	13			
RBC(2)	A≻B//C≻D	23	0	14	1	44,2	51,2	0,0771
	A≺B//C≺D	3	18	4	23			

(1) In group 5 (List), due to inconsistencies, it was not possible to obtain the indifference interval in 9 occasions for pair A-B and 5 cases for pair C-D. (2) In group 6 (RBC) the indifference interval could not be identified in 6 occasions for pair A,B and in 8 cases for pair C-D. (3) McNemar's test 2-sided.

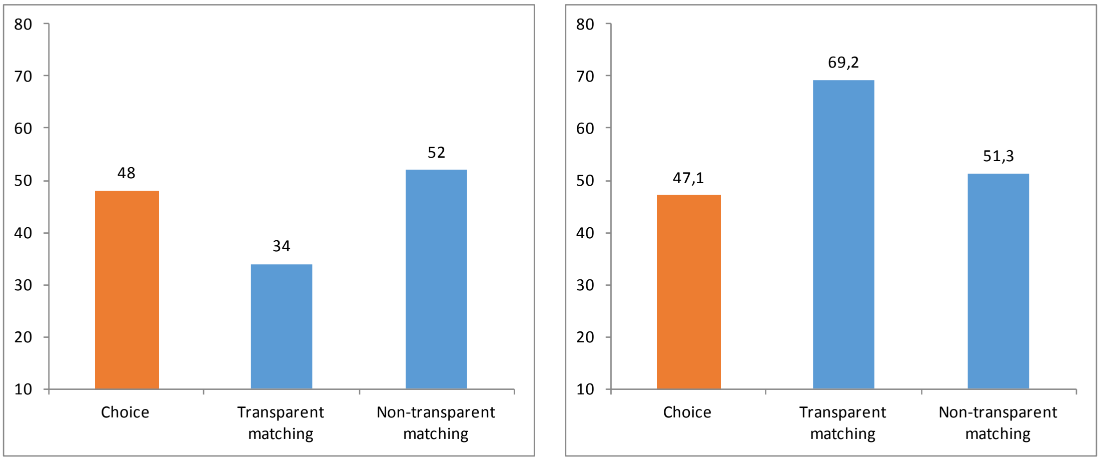
Table 5. Direct choices vs choices implied by the valuation task. Round 3.

		Valuation				A≻B//C≻D (%)		
	Choice	A≻B	A≺B	C≻D	C≺D	Choice	Valuation	p-value(3)
Bisection	A≻B//C≻D	30	0	20	0	50,0	68,0	<0,0001
	A≺B//C≺D	8	12	10	20			
Ping-pong	A≻B//C≻D	31	1	23	0	55,0	71,0	0,0004
	A≺B//C≺D	7	11	10	17			
HCBM	A≻B//C≻D	29	2	12	3	46,0	50,0	0,4227
	A≺B//C≺D	2	17	7	28			
List(1)	A≻B//C≻D	25	0	15	0	50,0	73,8	<0,0001
	A≺B//C≺D	8	9	11	12			
RBC(2)	A≻B//C≻D	25	1	17	0	51,8	57,8	0,1306
	A≺B//C≺D	3	14	3	20			

(1) In group 5 (List), due to inconsistencies, it was not possible to obtain the indifference interval in 8 occasions for pair A-B and 12 cases for pair C-D. (2) In group 6 (RBC) the indifference interval could not be identified in 7 occasions for pair A,B and in 10 cases for pair C-D. (3) McNemar's test 2-sided.

These results seem to support the explanation of Fischer et al (1999) about the origin of Preference Reversals between Choice and Matching. However, there are some differences between their results and ours that we believe are important in order to interpret the normative status of non-transparent methods. Figure 5 compares the results in Fischer et al (1999, Study 2), with our results. In our case, we assume the prominent attribute is the probability of the best outcome, as it is usually done in the literature about Preference Reversals in gambles (Slovic et al., 1990). There is one common result between our results and those in Fischer et al, namely, both studies show that when CBM is conducted using non-transparent methods, the disparity between Choice and Matching disappears. However, they seem to attribute this result to the Prominence effect, equally present in Choice and non-transparent Matching. That is, both methods are biased because they both overweight the prominent attribute. This conclusion is supported because in their study (as in ours) transparent matching was conducted using the Prominent attribute to equate both options. Assuming that Scale Compatibility favors the option that is better on the matching attribute, transparent matching was already overweighting the more prominent attribute. However, in Choice and in non-transparent matching, preferences for the option that was better on the Prominent attribute were even higher than in transparent matching. The implication of that is that Choice and non-transparent matching greatly overvalue the option that is better on the more prominent attribute. That is why the Choice-Matching disparity is avoided. However, our results suggest a different interpretation.

Figure 5. Percentage of preferences for the alternative that was superior on the prominent attribute (Salary in Fischer et al.'s Study 2; Probability in our Experiment).



Fischer et al. (1999) – Study 2; Figure 2 (p. 1066).

Aggregate results in our study (3 rounds)

We observe that preferences for the option that is better on the most prominent attribute (the P-bet) is more preferred in transparent than in non-transparent matching and Choice. We interpreted this result as produced from strong Scale Compatibility effects in transparent matching. In fact, there is evidence (Delquié, 1993; Bleichrodt and Pinto, 2002) that Scale Compatibility is present in CBM transparent methods. It seems that non-transparent methods are less influenced by Scale Compatibility than transparent methods, that is, they correct a bias. To what extent they are still biased by the Prominence effect is something we do not know.

4. CONCLUSIONS

The main objective of this paper has been to test which features of matching methods based on choices could reduce the disparity between Matching and Choice. We conclude that methods based on transparent iterations overweight the attribute used to establish indifference. There is not too much evidence of this effect in the health economics literature so this is the first contribution of this paper. The second lesson of this paper is: do not present all choices at the same time in a "list". For example, the widely used Holt-Laury procedure (Holt and Laury, 2002) to elicit preferences for risk or, more generally, the Multiple Choice List mechanism, has some resemblance with our List method. It seems that presenting all choices together increases the attraction of the attribute that changes in the process. The more positive result comes from non-transparent methods. Randomizing the choices (RBC) and the HCBM method reduced the disparity between valuation and choice. This result coincides with the results in Fischer et al (1999) and supports the Task-Goal hypothesis. We also find that non-transparent methods are less biased by Scale Compatibility.

What are the implications of these results for preference elicitation methods in health? They seem to be that iterative CBM methods should try to hide, as much as possible, the goal of the task. In that way, the subject seems to treat each choice in the iteration process independently from the rest, without being influenced by past choices or other considerations. Furthermore, the good results of HCBM suggest that it may not be necessary to abandon iterative methods and move to non-iterative ones like RBC (that is, Discrete Choice Experiments) to avoid Compatibility effects. Our results highlight the importance of the distinction between transparent and non-transparent methods. Although this distinction is not new, it has received little attention in the literature. In fact, it is not uncommon to find papers (see Bleichrodt et al., 2007) where researchers (ourselves) refer to choice-based methods as having some advantages over standard matching without making any reference to the transparency issue. However, our result and Fischer et al (1999) result support the conclusion that CBM does not avoid the problems of standard matching unless it is made non-transparent. In conclusion, if we want to use CBM methods to estimate utilities for health states non-transparent methods seem to reduce biases present in transparent methods, making choice and matching more similar. To what extent those preferences are still affected by other biases is something that is left to further research.

REFERENCES

Abellán, J.M., Sánchez, F.I., Martínez, J.E., Méndez, I. (2012). Lowering the Floor of the SF-6D algorithm using a lottery equivalent method. *Health Economics, 21*, 1271-1285.

Badia, X., Roset, R., Herdman, M., Kind, P. (2001). A comparison of GB and Spanish general population time trade-off values for EQ-5D health states. *Medical Decision Making 21*, 7-16.

Bansback, N., Brazier, J., Tsuchiya, A., Anis, A. (2012). "Using a Discrete Choice Experiment to Estimate Health State Utility Values." *J Health Econ* 31(1), 306-318.

Bleichrodt, H., Pinto, J. L. (2002). Loss aversion and scale compatibility in two-attribute trade-offs. *Journal of Mathematical Psychology*, *46*(3), 315-337.

Bleichrodt, H., Abellan-Perpiñan, J.M., Pinto-Prades, J.L., Mendez-Martinez, I. (2007). Resolving inconsistencies in utility measurement under risk: Tests of generalizations of expected utility. *Management Science*, *53*(3), 469-482.

Butler, D.J., Loomes, G.C. (2007). Imprecision as an account of the preference reversal phenomenon. *American Economic Review*, *97*(1), 277-297.

Clark, M.D., Determann, D., Petrou, S., Moro, D., de Beker-Grob, E.W. (2014). Discrete choice experiments in health economics: a review of the literature. *Pharmacoeconomics, 32*(9), 883-902.

Delquié, P. (1993). Inconsistent trade-offs between attributes: New evidence in preference assessment biases. *Management Science*, *39*(11), 1382-1395

Fischer G.W., Hawkins, S.A. (1993). Strategy Compatibility, scale Compatibility and the Prominence Efect. *Journal of Experimental Psychology: Human Perception and Performance, 19*(3), 580-597.

Fischer, G.W., Carmon, Z., Ariely, D., Zauberman, G. (1999). Goal-based construction of preferences: Task goals and the prominence effect. *Management Science*, *45*(8), 1057-1075.

Fitts, P.M., Seeger, C.M. (1953). S-R compatibility: Spatial characteristics of stimulus and response codes. *Journal of Experimental Psychology, 46*, 199-210

Holt, C.A., Laury, S.K. (2002). Risk Aversion and Incentive Effects. *The American Economic Review 92*(5), 1644-1655.

Lichtenstein, S., Slovic, P. (1971). "Reversals of preference between bids and choices in gambling decisions." *Journal of Experimental Psychology*, 89(1), 46-55.

McCord, M. de Neufville, R. (1986). Lottery equivalents: reduction of the certainty effect problem in utility assessment. *Management Science, 32*(1), 56-60.

Slovic, P., Griffin, D., Tversky, A. (1990). Compatibility effects in judgment and choice. In Robin M. Hogarth (ed.) *Insights in Decision Making: A Tribute to Hillel J. Einhorn*, Chicago: The University of Chicago Press, 5-27.

Sumner, W., Nease, F. (2001). Choice-Matching preference reversals in health outcome assessments. *Medical Decision Making* 21(3), 208-218.

APPENDIX

Results assuming that a preference is implied by the valuation task (PLE) when intervals do not overlap

Table A1. Direct choices vs choices implied by the valuation task Round 1.

		Valuation				A≻B//C≻D (%)			Undetermined preference in PLE[3]	
	Choice	A≻B	A≺B	C≻D	C≺D	Choice	Valuation	p-value[4]	A vs. B	C vs. D
Bisection	A≻B//C≻D	23	0	15	1	51,3	68,4	0,0019	5	0
	A≺B//C≺D	4	8	10	15				10	9
Ping-pong	A≻B//C≻D	26	2	18	0	53,5	70,9	0,0013	2	1
	A≺B//C≺D	6	8	11	15				6	5
HCBM	A≻B//C≻D	24	0	10	2	46,8	54,5	0,1138	5	1
	A≺B//C≺D	3	11	5	22				7	10
List[1]	A≻B//C≻D	16	0	10	0	35,6	78,1	0,0000	1	0
	A≺B//C≺D	12	9	19	7				6	8
RBC[2]	A≻B//C≻D	11	0	7	2	41,7	43,8	1,0000	1	0
	A≺B//C≺D	3	11	0	14				3	7

[1] In group 5 (List), due to inconsistencies, it was not possible to obtain the indifference interval in 6 occasions for pairs A-B and C-D. [2] In group 6 (RBC) the indifference interval could not be identified in 21 occasions for pair A-B and in 20 cases for pair C-D. [3] Cases in which preference cannot be inferred from the PLE responses, since indifference intervals for each lottery overlap. [4] McNemar's test 2-sided.

Table A2. Direct choices vs choices implied by the valuation task. Round 2.

		Valuation				A≻B//C≻D (%)			Undetermined preference in PLE[3]	
	Choice	A≻B	A≺B	C≻D	C≺D	Choice	Valuation	p-value[4]	A vs. B	C vs. D
Bisection	A≻B//C≻D	27	0	18	0	55,6	67,9	0,0044	3	1
	A≺B//C≺D	5	13	5	13				2	13
Ping-pong	A≻B//C≻D	34	1	22	0	63,3	72,2	0,0269	1	0
	A≺B//C≺D	0	8	9	16				6	3
HCBM	A≻B//C≻D	27	2	12	1	51,2	52,4	1,0000	1	3
	A≺B//C≺D	3	13	1	23				4	10
List[1]	A≻B//C≻D	26	0	17	0	57,3	78,7	0,0002	0	0
	A≺B//C≺D	5	7	11	9				3	8
RBC[2]	A≻B//C≻D	23	0	14	1	52,8	52,8	0,4795	0	0
	A≺B//C≺D	1	16	0	17				4	10

[1] In group 5 (List), due to inconsistencies, it was not possible to obtain the indifference interval in 9 occasions for pair A-B and in 5 cases for pair C-D. [2] In group 6 (RBC) the indifference interval could not be identified in 6 occasions for pair A-B and in 8 cases for pair C-D. [3] Cases in which preference cannot be inferred from the PLE responses, since indifference intervals for each lottery overlap. [4] McNemar's test 2-sided.

Table A3. Direct choices vs choices implied by the valuation task. Round 3.

		Valuation				A≻B//C≻D (%)			Undetermined preference in PLE[3]	
	Choice	A≻B	A≺B	C≻D	C≺D	Choice	Valuation	p-value[4]	A vs. B	C vs. D

Bisection	A≻B//C≻D	29	0	20	0	62,0	73,4	0,0077	1	0
	A≺B//C≺D	2	9	7	12				9	11
Ping-pong	A≻B//C≻D	29	0	23	0	60,5	74,4	0,0015	3	0
	A≺B//C≺D	5	7	7	15				6	5
HCBM	A≻B//C≻D	27	2	12	2	48,9	48,9	0,7237	2	1
	A≺B//C≺D	0	17	4	24				2	7
List[1]	A≻B//C≻D	25	0	14	0	54,9	77,5	0,0002	0	1
	A≺B//C≺D	7	7	9	9				3	5
RBC[2]	A≻B//C≻D	25	1	17	0	62,3	60,9	1,0000	0	0
	A≺B//C≺D	0	12	0	14				5	9

[1] In group 5 (List), due to inconsistencies, it was not possible to obtain the indifference interval in 8 occasions for pair A-B and in 12 cases for pair C-D. [2] In group 6 (RBC) the indifference interval could not be identified in 7 occasions for pair A-B and in 10 cases for pair C-D. [3] Cases in which preference cannot be inferred from the PLE responses, since indifference intervals for each lottery overlap. [4] McNemar's test 2-sided.