Theory and Decision (2006) 61: 51–62
DOI 10.1007/s11238-006-7191-7

ILIA TSETLIN

A METHOD FOR ELICITING UTILITIES AND ITS APPLICATION TO COLLECTIVE CHOICE

ABSTRACT. Designing a mechanism that provides a direct incentive for an individual to report her utility function over several alternatives is a difficult task. A framework for such mechanism design is the following: an individual (a decision maker) is faced with an optimization problem (e.g., maximization of expected utility), and a mechanism designer observes the decision maker's action. The mechanism does reveal the individual's utility truthfully if the mechanism designer, having observed the decision maker's action, infers the decision maker's utilities over several alternatives. This paper studies an example of such a mechanism and discusses its application to the problem of optimal social choice. Under certain simplifying assumptions about individuals' utility functions and about how voters choose their voting strategies, this mechanism selects the alternative that maximizes Harsanyi's social utility function and is Pareto-efficient.

KEY WORDS: decision making, social choice, uncertainty, utility elicitation

1. INTRODUCTION

Designing a mechanism that provides incentives for a decision maker to report her preferences truthfully is an important problem by itself. The key is to relate both the final payoff and the probabilities of possible outcomes to the decision maker's reported valuations of alternatives. Similar "honesty" problems occur in experts' evaluations of probabilities (Winkler, 1996) and bargaining procedures (Brams and Kilgour, 1996). This paper presents a setting where an individual, in order to maximize her expected utility, reveals her cardinal utility function over several available alternatives. The focus of the proposed mechanism is on providing incentive to reveal the utility truthfully, and not on helping the individual

to assess different complicated aspects of her utility function (Dyer and Sarin, 1982). Throughout the paper the individual is assumed to be risk neutral with constant marginal utility for money. These assumptions might be viewed as first-order approximations of a more complicated utility function.

The proposed mechanism of eliciting utilities for several alternatives works as follows: for each alternative, the decision maker chooses the sum of money which she pays (or receives) if this alternative occurs. The probability of each alternative increases with the increase of the corresponding sum. The only constraint is that the sum of money over all alternatives is zero. Thus the decision maker makes many tradeoffs, in particular between increasing the probability of the most preferred alternative and paying too much if this alternative occurs. The mechanism designer observes the decision maker's choice of the sums of money and then infers the utilities for all alternatives.

As soon as such a mechanism of revealing individuals' utilities exists, it is tempting to apply it to social choice problems. Since individuals do reveal their utilities, one can think about the social choice that maximizes Harsanyi's social utility function (Harsanyi, 1955). The implications of the above mechanism for social choice are discussed. Dyer and Sarin (1979) and Harvey (1999) propose group preference aggregation rules based on measurable value functions and not on utility functions. Since I assume that individuals are risk neutral, I do not make that distinction.

The next section formulates the problem so that the decision maker's solution to it reveals her utilities for several alternatives. Section 3 applies this result to the collective choice. Section 4 concludes.

2. ELICITING UTILITIES

Consider a decision maker facing the following problem: There are n alternatives $\{c_1, \ldots, c_n\}$. The occurrence of a particular alternative depends upon the realization of n

independent identically distributed random variables X_i, $i = 1, \ldots, n$, measured in units of money. If alternative c_i occurs, the decision maker's utility is $U(c_i, w)$, where w is her wealth.

The decision maker can influence the probabilities of the occurrence of the alternatives. She chooses n values (in monetary units) $\{s_1, \ldots, s_n\}$, subject to the constraint

$$\sum_{i=1}^{n} s_i = 0. \tag{1}$$

Then n independent random variables X_i are independently drawn from the distribution with differentiable probability density function (hereafter, pdf) $f(x)$ and alternative c_i with the largest value $x_i + s_i$ occurs. Then the decision maker pays the sum of money s_i (if s_i is negative, she receives the sum of money $-s_i$) and, therefore, her utility is $U(c_i, w - s_i)$. Denote by $P_i(s_1, \ldots, s_n)$ the probability of alternative c_i given the decision maker's strategy $\{s_1, \ldots, s_n\}$. Note that $P_i(0, \ldots, 0) = \frac{1}{n}$, since all X_i are independently drawn from the same distribution. The decision maker's goal is to maximize the expected utility $EU(s_1, \ldots, s_n)$ given by

$$EU(s_1, \ldots, s_n) = \sum_{i=1}^{n} U(c_i, w - s_i) P_i(s_1, \ldots, s_n). \tag{2}$$

In general, the choice of $\{s_1, \ldots, s_n\}$ that maximizes (2) subject to the constraint (1) depends upon risk attitude, marginal utility for money (that might be different for different c_i), and the functions $P_i(s_1, \ldots, s_n)$, $i = 1, \ldots, n$. For example, if $P_i(s_1, \ldots, s_n) = \frac{1}{n}$ for all s_i, then the maximization of (2) subject to the constraint (1) corresponds to the standard actuarially fair insurance problem. The focus of this section is on revealing the utilities $U(c_i, w)$ for alternatives c_i. If the assumptions below hold, then the choice of s_i reveals these utilities.

ASSUMPTION 2.1. The marginal utility for money does not depend upon alternative c_i and upon wealth w:

$$U(c_i, w - s_i) = U(c_i) + w - s_i.$$

ASSUMPTION 2.2. The decision maker's impact (by the choice of $s_1, \ldots, s_n$) on the probabilities $P_i(s_1, \ldots, s_n)$ is small. More formally, if $max_{i,j=1,\ldots,n}(|s_i - s_j|) = b$ then

$$P_i(s_1, \ldots, s_n) = \frac{1}{n} + \alpha \sum_{j \neq i} (s_i - s_j) + \alpha o(b),$$
$$\text{where } \alpha > 0, \ \lim_{\epsilon \to 0} \frac{o(\epsilon)}{\epsilon} = 0. \tag{3}$$

Assumption 2.2 without the term $\alpha o(b)$ corresponds to Merrill's (1981) framework. Alternatively, it is a first-order Taylor expansion of $P_i(s_1, \ldots, s_n)$ with α being the first derivative, and is valid if the variance of the distribution for X_i is large enough.

THEOREM 2.3. *Under Assumptions* 2.1, 2.2, *and constraint* (1), *the expected utility* (2) *is maximized at* $s_i = \frac{U(c_i) - \overline{U}}{2} + o(b)$, *where* $\overline{U} = \frac{1}{n} \sum_{i=1}^{n} U(c_i)$ *and* $b = max_{i,j=1,\ldots,n}(|U(c_i) - U(c_j)|)$.

Proof. Since $\sum_{i=1}^{n} P_i(s_1, \ldots, s_n) = 1$, the decision maker's expected utility (2) is

$$EU(s_1, \ldots, s_n) = w + \sum_{i=1}^{n} (U(c_i) - s_i) \times \left(\frac{1}{n} + \alpha \sum_{j \neq i} (s_i - s_j) + \alpha o(b) \right). \tag{4}$$

Note that constraint (1) implies $\sum_{j \neq i} (s_i - s_j) = n s_i$. Therefore, (4) becomes

$$EU(s_1, \ldots, s_n) = w + \sum_{i=1}^{n} (U(c_i) - s_i) \left(\frac{1}{n} + n\alpha s_i + \alpha o(b) \right)$$
$$= w + \overline{U} + n\alpha \sum_{i=1}^{n} (U(c_i) s_i - s_i^2) + n\alpha \overline{U} o(b).$$

Except for $n\alpha \overline{U} o(b)$, the only term that depends upon s_i is $n\alpha \sum_{i=1}^{n} (U(c_i) s_i - s_i^2)$, which can be rewritten as

$$\sum_{i=1}^{n}(U(c_i)s_i-s_i^2)=-\sum_{i=1}^{n}\left(s_i-\frac{U(c_i)-\overline{U}}{2}\right)^2+\sum_{i=1}^{n}\left(\frac{U(c_i)-\overline{U}}{2}\right)^2.$$

This term is maximized at $s_i^*=\frac{U(c_i)-\overline{U}}{2}$, $i=1,\ldots,n$, which satisfy constraint (1). Therefore, (4) is maximized at $s_i=s_i^*+o(b)=\frac{U(c_i)-\overline{U}}{2}+o(b)$. □

Example. Consider the case of $n=2$ alternatives. Random variables X_1 and X_2 are independently identically distributed. Denote $X=X_1-X_2$, so X has a pdf $f_{12}(x)$ which is symmetric about zero. Denote the corresponding cumulative distribution function by $F_{12}(x)$. Note that $F_{12}(0)=\frac{1}{2}$. In that notation,

$$\begin{aligned}P_1(s_1,s_2)&=P(X_1+s_1>X_2+s_2)=P(X>s_2-s_1)\\&=1-F_{12}(s_2-s_1).\end{aligned}$$

The first-order Taylor expansion yields

$$P_1(s_1,s_2)=\frac{1}{2}+f_{12}(0)(s_1-s_2)+o(|s_2-s_1|),$$

which corresponds to (3). By Theorem 2.3, the decision maker chooses $s_1=\frac{U(c_1)-U(c_2)}{4}+o(U(c_1)-U(c_2))$ and $s_2=\frac{U(c_2)-U(c_1)}{4}+o(U(c_1)-U(c_2))$. Note that the choice of s_i does not depend upon the distribution of X_i.

3. IMPLICATION FOR SOCIAL CHOICE

Consider a social choice problem: among n alternatives, choosing one that, by some criteria, is the best for the society. Definition 3.1 and Theorem 3.2 below characterize social choice that is Pareto-efficient when allowing for wealth redistribution. Let the number of individuals in the society be m, and let individual j's utility for alternative c_i and wealth w_j be $U_j(c_i,w_j)$, $j=1,\ldots,m$.

DEFINITION 3.1. The social choice c_i is Pareto-efficient when allowing for wealth redistribution, if for any c_k and any δ_j, with

$k=1,\ldots,n$, $j=1,\ldots,m$, and $\sum_{j=1}^{m}\delta_j=0$, there exists individual j_0 such that $U_{j_0}(c_i, w_{j_0}) > U_{j_0}(c_k, w_{j_0}+\delta_{j_0})$.

In words, the social choice is efficient if it cannot be Pareto-improved by choosing another alternative and redistributing wealth (making transfer payments).

THEOREM 3.2. *Let individual j's utility have a form $U_j(c_i, w_j) = U_j(c_i) + w_j$ for all j (i.e., let Assumption 2.1 hold). Then the only social choice c_i which is Pareto-efficient when allowing for wealth redistribution, is the one that maximizes*

$$W(c_i) = \sum_{j=1}^{m} U_j(c_i). \tag{5}$$

Proof. Consider c_i such that $\sum_{j=1}^{m} U_j(c_i) > \sum_{j=1}^{m} U_j(c_k)\ \forall k \neq i$. Suppose that social choice c_i is not efficient. Then, there exist c_k and δ_j, with $\sum_{j=1}^{m}\delta_j = 0$, such that $U_j(c_i, w_j) \leq U_j(c_k, w_j+\delta_j)\ \forall j$. By Assumption 2.1 this is equivalent to $U_j(c_i) + w_j \leq U_j(c_k) + w_j + \delta_j\ \forall j$. Taking the sum over j yields $\sum_{j=1}^{m} U_j(c_i) \leq \sum_{j=1}^{m} U_j(c_k) + \sum_{j=1}^{m}\delta_j$, a contradiction. □

Note that, under Assumption 2.1, maximization of $W(c_i)$, given by (5), corresponds to maximization of Harsanyi's utility function (Harsanyi, 1955).

The rest of this section applies the mechanism of revealing individual's preferences, described in Section 2 (Theorem 2.3), to the social choice that maximizes Harsanyi's utility function (5) and, by Theorem 3.2, leads to selecting the alternative that is Pareto-efficient when allowing for wealth redistribution.

Consider the following hypothetical voting system. There are n alternatives $\{c_1,\ldots,c_n\}$. Voter j, $j=1,\ldots,m$, chooses sums of money s_{ij} by which she votes for alternative c_i with the only constraint $\sum_{i=1}^{n} s_{ij} = 0$. After all voters have submitted their s_{ij}, alternative c_i receives the score $S_i = \sum_{j=1}^{m} s_{ij}$, and the alternative with the highest S_i is elected. Voter j pays the sum of money s_{ij} (if s_{ij} is negative, she receives $-s_{ij}$), and the remaining surplus, which equals S_i, is, e.g., either evenly

distributed among all the voters or used for some public needs. Note that this surplus is always positive since $\sum_{i=1}^{n} S_i = 0$ and the alternative with the highest S_i is elected. Furthermore, if the number of voters is high, voters' behavior does not depend on how this surplus is divided.

The outcome of such an election depends upon voters' assumptions about others' voting strategies. If every voter assumes that the number of voters is large and all candidates are equally likely to win, then every voter votes as assumed by Theorem 2.3.

Observation 3.3. If every voter chooses $s_{ij} = \frac{U_j(c_i) - \overline{U_j}}{2}$, where $\overline{U_j} = \frac{1}{n}\sum_{i=1}^{n} U(c_i)$, then alternative c_i that maximizes Harsanyi's utility function (5) is elected.

Proof. $S_i = \sum_{j=1}^{m} s_{ij} = \sum_{j=1}^{m} \left(\frac{U_j(c_i) - \bar{U}_j}{2}\right) = \frac{1}{2}\sum_{j=1}^{m} U_j(c_i) - \frac{1}{2}\sum_{j=1}^{m} \overline{U_j}$. Therefore,

$$S_i > S_k \Leftrightarrow \sum_{j=1}^{m} U_j(c_i) > \sum_{j=1}^{m} U_j(c_k). \qquad \square$$

Thus, if voters' beliefs correspond to the framework of Section 2, then, by Theorem 2.3, they would vote according to Observation 3.3, and thus, by Theorem 3.2, the social choice will be Pareto-efficient when allowing for wealth redistribution. The remainder of this section discusses and tries to justify the assumption that voters might indeed vote as assumed above, especially in a large electorate.

Every voter j faces exactly the same optimization problem as described in Section 3, with random variables X_i corresponding to aggregated scores for each alternative, excluding voter's own votes s_{ij}. Therefore, if her beliefs are that X_i are independently drawn from a distribution with large variance, the assumptions of Theorem 2.3 apply. For a large electorate the variance of X_i is large, compared to the individual's impact.

It is important that the probability distribution over X_i represents a voter's belief about other voters' actions, not about their preferences. In this framework, Gibbard's (1973) theorem can be

interpreted as "there is no single strategy that a voter should follow independent of her beliefs." However, the voter still should choose some strategy, and for this she should make assumptions about other voters' strategies. This can lead to completely intractable and/or counter-practical conclusions, as Rubinstein (1989) shows in the elegant example about fragility of the common knowledge assumption. However, whatever infinite regression is made, a voter ends with some "working hypothesis" about the joint distribution of X_i, and makes her choice according to this working distribution. I would like to step away from game-theoretical complications of how this distribution is constructed, whether it corresponds to Bayesian Nash Equilibrium (which might not exist for this game) etc., and just argue that the assumptions of Theorem 2.3 might hold.[1]

For example, suppose that there are two candidates running for election, and common knowledge is that 90% of the voters prefer c_1 to c_2. If voter j assumes that everybody is going to vote sincerely, then she herself would vote $s_{1j} = -M$ and $s_{2j} = M$, where $M >> 0$ is very large. This voting strategy does not depend on whether (or "how strongly") voter j prefers c_1 or c_2. Given this, voter j may assume that everybody else thinks the same way, and then alternative c_2 would be elected. Under this belief voter j should vote $s_{1j} = M$ and $s_{2j} = -M$. This regression can go forever, but finally every voter might assume that all X_i are independently drawn from some distribution $f(x)$ with large variance. Given this belief, everybody chooses $s_{ij} = \frac{U_j(c_i) - \overline{U_j}}{2}$ as described in Section 2.

A similar framework and approach to the analysis of voters' behavior is used by Merrill (1981). Merrill studies different voting procedures (procedures differ by restrictions on admissible s_i), but in his work the final payoff to a voter does not depend upon a voting strategy, i.e., instead of maximizing $E[U(c_i, w - s_i)]$, a voter maximizes $E[U(c_i)]$. As a consequence, these procedures do not completely reveal and aggregate individual voters' utilities.

4. CONCLUSIONS

Section 2 provides a mechanism for revealing individual's utilities for several alternatives under two assumptions. Assumption 2.1 abstracts away many interesting features of utility functions. However, if all alternatives c_i are close enough to each other, so that the effects of risk aversion and of alternative-dependent marginal utilities for money are small, this assumption might be viewed as a first-order approximation of a more complicated utility function. An attractive property of the proposed mechanism is that if the individual's impact on the probabilities of the alternatives is small (Assumption 2.2), then the choice of decision variables directly reveals the utilities for all alternatives (Theorem 2.3).

Section 3 discusses the implications of this utility-revealing mechanism for social choice. Theorem 3.2 states a particular form of Harsanyi's utility function, which, under Assumption 2.1, is Pareto-efficient when allowing for wealth redistribution. Observation 3.3, combined with Theorem 2.3, shows that such Pareto-efficient alternative is elected if voters' beliefs and utilities correspond to the setting of Section 2.

Gibbard (1973) have shown that there is no dominant strategy for an individual to *always* report her preferences (or utilities) truthfully. Given that result, researchers studied "reasonable" strategies under some assumptions about an individual's impact on the aggregated outcome. Merrill (1981) provides optimal voting strategies for different voting rules, such as approval voting, Borda score, and Z-score. In each of these methods, a voter casts a particular number of points in favor of each alternative, and all systems differ in the restrictions on the admissible points. None of those voting rules is able to elicit voters' preferences truthfully, which is the easiest to see in the example with two voters and two alternatives: one is almost indifferent between alternatives c_1 and c_2, but would prefer c_1, and the other voter strongly prefers c_2 to c_1. It seems that the only way to provide an incentive for individuals to reveal the strength of their preferences truthfully is to relate the voter's payoff, in the case of a particular alternative

being elected, to her voting strategy. Similar wealth redistribution, contingent upon the election outcome, might occur, e.g., via trading election-contingent securities (Musto and Yilmaz, 2003).

This paper provides one example of a voting system that elicits and aggregates individuals' utilities under the assumptions of a small individual impact on the final outcome, constant marginal utility for money and a symmetric probability distribution over candidates to be elected (Observation 3.3). Although the assumption about symmetry might seem highly artificial, some supporting arguments are provided at the end of Section 3. It is also worth noting that the assumption that all alternatives are equally likely to be elected is consistent with *ignorance priors* heuristic in individual decision making, that has strong experimental evidence (Fox and Rottenstreich, 2003; Rottenstreich and Tversky, 1997; Tversky and Kahneman, 1973, 1974; Tversky and Koehler, 1994).

In this paper, individuals' utilities are scaled in money. Though in many social choice problems comparison of individuals' utilities in monetary units is wrong and amoral, money is a convenient common denominator since it may be transferred from one individual to another. As a consequence, the social choice made according to the proposed procedure is Pareto-efficient with respect to switching to *any* other alternative and simultaneously making *any* transfer payments. Furthermore, sometimes a collective choice has a clear money-based content, as it is in a collective decision of shareholders. Measuring individuals' utilities in money seems to be the only reasonable approach to interpersonal comparison of utilities in this case. Social choice, made according to the proposed rule, has two attractive features: (1) it corresponds to the maximization of Harsanyi's social utility function and (2) the outcome is Pareto-efficient with respect to selecting any other alternative and simultaneously making any transfer payments. In particular, it might be adequate for a collective of shareholders' decision, especially if every shareholder possesses only a small stake of the company.

ACKNOWLEDGEMENTS

I thank Steven Brams, Alessandra Cillo, Enrico Diecidue, Robert Nau, Peter Wakker, and Robert Winkler for very helpful comments.

NOTES

1. The same problem about the choice of voting strategy and appropriate beliefs about other voters' actions arises for any voting rule, as guaranteed by Gibbard (1973). Taylor (2005) provides extensive discussion and many related results.

REFERENCES

Brams, S. and Kilgour, M. (1996), Bargaining procedures that induce honesty, *Group Decision and Negotiation* 5, 239–262.

Dyer, J. and Sarin, R. (1979), Group preference aggregation rules based on strength of preference, *Management Science* 25, 822–832.

Dyer, J. and Sarin, R. (1982), Relative risk aversion, *Management Science* 28, 875–886.

Fox, C. and Rottenstreich, Y. (2003), Partition priming in judgment under uncertainty, *Psychological Science* 14, 195–200.

Gibbard, A. (1973), Manipulation of voting schemes: a general result, *Econometrica* 41, 587–601.

Harsanyi, J. (1955), Cardinal welfare, individual ethics, and interpersonal comparability of utility, *Journal of Political Economy* 61, 309–321.

Harvey, C. (1999), Aggregation of individuals' preference intensities into social preference intensity, *Social Choice and Welfare* 16, 65–79.

Merrill, S. (1981), Strategic decisions under one-stage multi-candidate voting systems, *Public Choice* 36, 115–134.

Musto, D. and Yilmaz, B. (2003), Trading and voting, *Journal of Political Economy* 111, 990–1003.

Rottenstreich, Y. and Tversky, A. (1997), Unpacking, repacking, and anchoring: advances in support theory, *Psychological Review* 104, 406–415.

Rubinstein, A. (1989), The electronic mail game: strategic behavior under "almost common knowledge", *The American Economic Review* 79, 385–391.

Taylor, A. (2005), *Social Choice and the Mathematics of Manipulation*, Cambridge University Press.

Tversky, A. and Kahneman, D. (1973), Availability: a heuristic for judging frequency and probability, *Cognitive Psychology* 4, 207–232.
Tversky, A. and Kahneman, D. (1974), Judgment under uncertainty: heuristics and biases, *Science* 185, 1124–1131.
Tversky, A. and Koehler, D. (1994), Support theory: a nonextensional representation of subjective probability, *Psychological Review* 101, 547–567.
Winkler, R. (1996), Scoring rules and the evaluation of probabilities, *Test* 5, 1–60.

Address for correspondence: I. Tsetlin, INSEAD, 1 Ayer Rajah Avenue, 138676 Singapore, Singapore. E-mail: ilia.tsetlin@insead.edu