99-25

UNIVERSITY OF CALIFORNIA, SAN DIEGO

DEPARTMENT OF ECONOMICS

A PEDAGOGICAL PROOF OF ARROW'S IMPOSSIBILITY THEOREM

BY

VALENTINO DARDANONI

DISCUSSION PAPER 99-25
DECEMBER 1999

A PEDAGOGICAL PROOF OF ARROW'S IMPOSSIBILITY THEOREM

VALENTINO DARDANONI, UNIVERSITÀ DI PALERMO

ABSTRACT

In this note I consider a simple proof of Arrow's Impossibility Theorem (Arrow 1963). I start with the case of three individuals who have preferences on three alternatives. In this special case there are $13^3 = 2197$ possible combinations of the three individuals' rational preferences. However, by considering the subset of *linear* preferences, and employing the full strength of the IIA axiom, I reduce the number of cases necessary to completely describe the SWF to a small number, allowing an elementary proof suitable for most undergraduate students.

This special case conveys the nature of Arrow's result. It is well known that the restriction to three options is not really limiting (any larger set of alternatives can be broken down into triplets, and any inconsistency within a triplet implies an inconsistency on the larger set). However, the general case of $n \geq 3$ individuals can be easily considered in this framework, by building on the proof of the simpler case. I hope that a motivated student, having mastered the simple case of three individuals, will find this extension approachable and rewarding.

This approach can be compared with the traditional simple proofs of Barberà (1980), Blau (1972), Denicolò (1996), Fishburn (1970), Kelly (1988), Mueller (1989), Riker and Ordeshook (1973), Sen (1979,1986), Suzumura (1988) and Taylor (1995).

Suppose we have 3 individuals, a, b and c, who have preferences on 3 alternatives, x, y and z. Denote weak preference by $\succeq_i$, strict preference by $\succ_i$ and indifference by $\sim_i$, $i = a, b, c, s$, with s denoting society. A rational preference relation is complete and transitive (a weak ordering). A preference relation is *linear* when it is also antisymmetric, so that no two distinct alternatives are ever indifferent. Note that preferences can always be described by considering the ranking in the three *pairwise comparisons* x versus y, y versus z and z versus x.

Definitions: *A Rational Unrestricted-Domain Social Welfare Function (SWF) is a function that takes any three rational individual preferences and gives back a rational social preference. A SWF satisfies Independence of Irrelevant Alternatives (IIA) when social ranking on a given pairwise comparison depends only on individuals' ranking on that comparison. A SWF satisfies Unanimity (U) if, whenever everybody has the same strict ranking on a given pairwise comparison, then society has the same ranking. Individual i is a dictator if society's preference always coincide with individual's i strict preference regardless of all the other individuals' preferences. A SWF satisfies Non-Dictatorship (ND) if there is no dictator.*

Date: September 1999.

I would like to thank Massimo Bordignon, Vincenzo Denicolò, Walter Heller, Marco LiCalzi, Salvatore Modica, seminar audiences at UC San Diego and Pavia, students at UC San Diego and Palermo, and especially Mark Machina and two anonymous referees for very helpful comments on this material. Thanks to the Department of Economics at UC San Diego for hospitality and to MURST for financial assistance. Address for correspondence: Istituto di Scienze Finanziarie, Facoltà di Economia, Università di Palermo, Viale delle Scienze, 90132 Palermo, Italy. email: vdardano@unipa.it.

Arrow's Impossibility Theorem: *There is no SWF which satisfies U, IIA and ND.*

PROOF: The proof consists of two steps: in the first, I prove that if there is a disagreement on any given pairwise comparison, the SWF must agree with the majority. In the second step, I prove that this property implies the intransitivity of the SWF. Thus, the two steps jointly show that the axioms are inconsistent.

To prove the Theorem, extensive use will be made of the following table:

	x versus y								y versus z								z versus x							
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
a	$\succ$	$\succ$	$\succ$	$\succ$	$\prec$	$\prec$	$\prec$	$\prec$	$\succ$	$\succ$	$\succ$	$\succ$	$\prec$	$\prec$	$\prec$	$\prec$	$\succ$	$\succ$	$\succ$	$\succ$	$\prec$	$\prec$	$\prec$	$\prec$
b	$\succ$	$\succ$	$\prec$	$\prec$	$\succ$	$\succ$	$\prec$	$\prec$	$\succ$	$\succ$	$\prec$	$\prec$	$\succ$	$\succ$	$\prec$	$\prec$	$\succ$	$\succ$	$\prec$	$\prec$	$\succ$	$\succ$	$\prec$	$\prec$
c	$\succ$	$\prec$	$\succ$	$\prec$	$\succ$	$\prec$	$\succ$	$\prec$	$\succ$	$\prec$	$\succ$	$\prec$	$\succ$	$\prec$	$\succ$	$\prec$	$\succ$	$\prec$	$\succ$	$\prec$	$\succ$	$\prec$	$\succ$	$\prec$
s	$\succ$							$\prec$	$\succ$							$\prec$	$\succ$							$\prec$
s		$\prec$	$\succ$	$\prec$	$\succ$	$\prec$	$\succ$			$\prec$	$\succ$	$\prec$	$\succ$	$\prec$	$\succ$			$\prec$	$\succ$	$\prec$	$\succ$	$\prec$	$\succ$	

Table 1

where each column in the table describes a particular ranking for the three individuals in a given pairwise comparison: the first eight columns refer to x versus y, columns 9 to 16 refer to y versus z, the last eight columns refer to z versus x. The first three rows denote a's, b's and c's preferences, and the last two society's preferences. In the sequel, when I refer to a given column of the table, I will refer to a particular profile of the individuals' preference on a given comparison.

A few things are worth noting about this table: first, any given linear preference profile for the three individuals in this society can be described by appropriately picking one column from each pairwise comparison: for example, columns (5,11,18) describe $y \succ_a z \succ_a x$, $z \succ_b x \succ_b y$ and $x \succ_c y \succ_c z$. However, note that only specific combinations of columns will define *rational* preference profiles: for example, columns (1,9,23) violate the transitivity of c's preferences. It may be visually appealing to show students that the only intransitive linear preference profiles are those which result in $(\succ, \succ, \succ)$ or $(\prec, \prec, \prec)$. Second, note that, by IIA, knowing what social preference is under each of these 24 columns delivers a complete description of the SWF. Third, by Unrestricted Domain and Rationality, for any rational preference profile we decide to pick, there must correspond rational social preferences. Finally, note that some columns are already filled by invoking U. These are indicated in the fourth row of the table.

Step 1: Notice that in all the columns of the table which are not filled by U, we have a conflict between a majority of two individuals and a single dissenter. I first show that in the presence of such conflict, society cannot be indifferent. Secondly I show that if society sides with the single dissenting individual in a given case of disagreement, then this individual must be a dictator. Given that dictatorship is ruled out by axiom, this proves the step.

Take any case of disagreement of individual preferences on a given comparison, say, without loss of generality, column 2 in the table. Consider two alternative preference profiles: i) columns (2,9,23); and ii) columns (2,16,23). By IIA, social preference in the (x-y) and (z-x) comparisons must be the same in both cases. If in column 23 we have $x \sim_s z$, then $y \succ_s x \sim_s z$ in the first case and $z \sim_s x \succ_s y$ in the second. If $x \succ_s z$, transitivity forces $x \succ_s y$. If $z \succ_s x$, transitivity forces $y \succ_s x$. Therefore, in column 2 we must have either $x \succ_s y$ or $y \succ_s x$ (i.e. x cannot be socially indifferent to y).

Suppose then that $y \succ_s x$. Consider the preference profiles (2,16,19), (2,16,21) and (2,16,23). By assumption, $y \succ_s x$, and by U, $z \succ_s y$. Thus, by transitivity, $z \succ_s x$ in columns 19, 21 and 23 of row 5. Using the same reasoning, we choose the preference profiles: i) (2,11,24), (2,13,24), (2,15,24); ii) (1,15,18), (1,15,20), (1,15,22); iii) (1,10,23), (1,12,23), (1,14,23); iv) (3,16,18), (5,16,18), (7,16,18);

v) (4,9,23), (6,9,23) to fill the other entries in that row.[1] But rows 4 and 5 jointly imply that social preference is identical to row 3, that is, c is a dictator, and the step is proved.[2]

Step 2: Just take any 'voting paradox' preference profile, say (5,11,18), and use step 1 to get $x \succ_s y \succ_s z \succ_s x$, which violates transitivity. This concludes the proof of the Theorem for the case of three individuals.

I suspect that many teachers may wish to stop right here, feeling that they have already conveyed the essence of Arrow's argument. However, before turning to the general case of $n \geq 3$ individuals, we pause here to note that Arrow's Theorem works also for the simpler case of two individuals, by considering the subtable of table 1 made of columns (1,2,7,8,9,10,15,16,17,18,23,24). Individuals a and b have the same preferences in that subtable, and thus may be considered as a single individual, and a simplified step 1 suffices to prove the Theorem.[3]

It turns out that the above proof can be easily extended to the general case of $n \geq 3$ individuals. I will again proceed in two steps: in the first, I establish that *whenever there is a situation such that, on a given pairwise comparison, one single individual has dissenting preferences from the other $n-1$ individuals, the SWF must agree with the $n-1$ individuals.* The second step will show that *this leads to a contradiction.*

*Step 1**: Without loss of generality, assume that n wins on a pairwise comparison where she is unanimously opposed by all the other $n-1$ individuals: for example, assume that $y \succ_n x$, $x \succ_i y$, $i = 1, \cdots, n-1$, and $y \succ_s x$. Consider then an arbitrary profile of preferences on a given arbitrary pairwise comparison. Given individual n, partition the remaining $n-1$ individuals into two groups, call them A and B, such that everybody in A has the same preference as the nth individual, and everybody in B has a dissenting preference.[4] Let individual n play the role of c, individuals in A play the role of a, and individuals in B play the role of b in the proof of step 1 of the previous section. Then by translating the chosen arbitrary profile into table 1 above, and following the appropriate sequence of preference profiles, it follows that if she wins under column 2, she must also win in this arbitrary case. This clearly contradicts ND, and the step is proved.

*Step 2**: Consider the following sequence of preference profiles:

[1] Alternatively, row 5 could be filled in two steps: the first to demonstrate that if c wins on a given pairwise comparison when she is unanimously opposed by the other two individuals, she will always win in all instances where she is unanimously opposed (e.g. using profiles (2,16,23), (2,15,24), (1,15,18), (1,10,23), (7,16,18)); the second to demonstrate that additional support by another individual does not reduce her power (e.g. using profiles (2,16,19), (2,16,21), (2,11,24), (2,13,24), (1,15,20), (1,15,22), (1,12,23), (1,14,23), (3,16,18), (5,16,18), (4,9,23), (6,9,23)). In this second step, it may be worth noting that each time we enter a new social preference in row 5, a's and b's rational preferences in that column could be changed at will, including the possibility of indifference. Thus this is where it is convenient to relax the linearity assumption, if desired.

[2] When individual preferences are not linear, the proof works for this definition of dictatorship, where a dictator enforces *strict* preferences. A stronger definition of dictatorship, requiring that social preferences be identical to that of the dictator (including indifferences), would invalidate the proof of the Theorem. For example, this stronger definition of dictatorship in conjunction with the other axioms is compatible with the following SWF: on each pairwise comparison social preference agrees with c when she expresses a strict preference, but agrees with a whenever c is indifferent.

[3] In fact, depending on the level of the class, some teachers may want to simply use the table with only two individuals and 12 columns, not bothering to mention the case of three or more individuals. In my experience, the resulting proof is typically understood by most (if not all) undergraduate students in less than one hour of lecture.

[4] If it happens that the remaining $n-1$ individuals all agree with each other on this comparison, then partition them arbitrarily.

1	$\succ$	$\succ$	$\prec$	$\prec$	$\succ$	$\prec$	$\cdots$	$\cdots$	$\succ$	$\prec$	$\prec$
2	$\succ$	$\succ$	$\prec$	$\prec$	$\succ$	$\prec$	$\cdots$	$\cdots$	$\prec$	$\succ$	$\prec$
3	$\succ$	$\succ$	$\prec$	$\prec$	$\succ$	$\prec$	$\cdots$	$\cdots$	$\prec$	$\prec$	$\succ$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\cdots$	$\vdots$	$\vdots$	$\vdots$
n-3	$\succ$	$\succ$	$\prec$	$\prec$	$\succ$	$\prec$	$\cdots$	$\cdots$	$\prec$	$\prec$	$\succ$
n-2	$\succ$	$\succ$	$\prec$	$\succ$	$\prec$	$\prec$	$\cdots$	$\cdots$	$\prec$	$\prec$	$\succ$
n-1	$\succ$	$\prec$	$\succ$	$\prec$	$\prec$	$\succ$	$\cdots$	$\cdots$	$\prec$	$\prec$	$\succ$
n	$\prec$	$\succ$	$\succ$	$\prec$	$\prec$	$\succ$	$\cdots$	$\cdots$	$\prec$	$\prec$	$\succ$

Table 2

Consider the first preference profile, defined by the first 3 columns in the table. By step 1*, $x \succ_s y$ and $y \succ_s z$. Thus, $x \succ_s z$. But this implies that in the (z-x) comparison the SWF agrees with the first $n-2$ individuals when they are opposed by the last 2. Consider then the next preference profile, defined by columns 4, 5 and 6 in the table. By IIA, $x \succ_s z$, and by step 1*, $y \succ_s x$. Thus, $y \succ_s z$, that is, in the (y-z) comparison the SWF agrees with the first $n-3$ individuals when they are opposed by the last 3. It is obvious then that we can continue in this fashion until we eventually arrive at a preference profile such as the one defined by the last 3 columns of the table, where on a given pairwise comparison, say the (z-x), the SWF agrees with the first 2 individuals when they are opposed by the last $n-2$. Thus, under this preference profile we have $x \succ_s z$ but $z \succ_s y$ and $y \succ_s x$ by step 1*, a contradiction of transitivity. This concludes the proof the Theorem.

The reader may have realized that the last step of this proof is actually based on an induction argument. According to the level of mathematical sophistication of the audience, this can be made explicit, or left implicit as above. Step 2* can be proven alternatively by using Barberà's (1980) *pivotal voter* idea: one individual is pivotal for a given pairwise comparison at a preference profile if she can change the social preference by changing her preference.

Consider the following sequence of preference profiles on a given pairwise comparison, say the (x-y)'s:

x versus y

1	$\succ$	$\succ$	$\succ$	$\cdots$	$\succ$	$\succ$	$\prec$
2	$\succ$	$\succ$	$\succ$	$\cdots$	$\succ$	$\prec$	$\prec$
3	$\succ$	$\succ$	$\succ$	$\cdots$	$\prec$	$\prec$	$\prec$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\vdots$
n-2	$\succ$	$\succ$	$\succ$	$\cdots$	$\prec$	$\prec$	$\prec$
n-1	$\succ$	$\succ$	$\prec$	$\cdots$	$\prec$	$\prec$	$\prec$
n	$\succ$	$\prec$	$\prec$	$\cdots$	$\prec$	$\prec$	$\prec$
s	$\succ$	$\succ$				$\prec$	$\prec$

Table 3

where the first and the last two elements of the last row are filled by invoking U and step 1*. Going from the first column to the last, consider the first occurrence where $y \succ_s x$. Note than that in the column immediately preceding it $x \succ_s y$, but only one individual has changed her preferences. Let this individual be the kth. Given individual k, partition the remaining $n-1$ individuals into two groups, call them A and B, such that A contains the first $k-1$ individuals and B contains the last $n-k$. Let individual k play the role of c, individuals in A play the role of a, and individuals in B play the role of b in table 1. Note that in column 3 we have $x \succ_s y$ while in column 4 $y \succ_s x$ because c is pivotal in this pairwise comparison. Consider then the preference profile (3,9,22). $x \succ_s y$ and $y \succ_s z$ imply $x \succ_s z$ in column 22. But using the preference profile (4,15,22) we get $y \succ_s x$ and $x \succ_s z$ but $y \succ_s z$ by step 1*, contradicting the transitivity of the SWF.

REFERENCES

Arrow K J (1963) Social choice and individual values. 2nd ed., J. Wiley: NY.

Barberà, S (1980) Pivotal voters. A new proof of Arrow's theorem. Econ Letters 6: 13-16.

Blau JH (1972) A direct proof of the Arrow's theorem. Econometrica 40: 61-67.

Denicolò V (1996) An elementary proof of Arrow's impossibility theorem. The Japanese Economic Review 47: 432-35.

Fishburn PC (1970) Arrow's impossibility theorem: Concise proof and infinite voters. Journal of Econ Theory 2: 103-106.

Kelly JS (1988) Social choice theory: An introduction. Springer: NY.

Mueller DC (1989) Public choice II. Cambridge University Press: Cambridge.

Riker WH and PC Ordeshook (1973) An introduction to positive political theory. Prentice-Hall: NY.

Sen AK (1979) Personal utilities and public judgements: Or what's wrong with welfare economics? The Economic Journal 89: 537-558.

Sen AK (1986) Social choice theory, in KJ Arrow and MD Intrilligator (eds.) Handbook of Mathematical Economics, vol.3, North Holland: Amsterdam.

Suzumura K (1988) Reduction of social choice problems: A simple proof of Arrow's general possibility theorem. Hitotsubashi Journal of Economics 29: 219-221.

Taylor AD (1995) Mathematics and politics. Springer-Verlag: NY.