Reprinted from JOURNAL OF ECONOMIC THEORY
All Rights Reserved by Academic Press, New York and London

Vol. 23, No. 2, October 1980
Printed in Belgium

Stable Voting Schemes

SALVADOR BARBERA

Departamento de Teoría Económica, Facultad de Ciencias Económicas, Universidad del Pais Vasco, Bilbao, Spain

Received December 6, 1978; revised September 24, 1979

I. INTRODUCTION

The Gibbard–Satterthwaite theorem establishes that no nondictatorial voting scheme can assign single-valued outcomes to preference profiles in such a way that correct revelation of preferences by all individuals is always a Nash-equilibrium of the games where preferences are the strategies, the voting scheme is the outcome function, and payoffs are determined by the true preferences of individuals. An interpretation of this result is that revealing one's own preferences is not always best. Therefore, it may be worthwhile for individuals to gather information on the workings of the voting scheme and on the voting attitudes of other members of society, and to act strategically on the basis of such information [2, 5].

Following this celebrated theorem, a number of works have explored the consequences of relaxing some of the somewhat restrictive assumptions on which it rests. Among the several contributions of Pattanaik to the subject is that of considering the effects of retaining notions of game-theoretic equilibrium other than Nash's, on the conclusion of Gibbard and Satterthwaite.

Indeed the Gibbard–Satterthwaite theorem proves that individuals will have an incentive to act strategically. But it does not necessarily imply that individuals will actually misrepresent their preferences whenever this appears to be profitable: their actual behavior will depend on a number of considerations which are simply beyond the scope of Nash-equilibrium analysis. One such factor is the quality of information available to individuals. Another, the one we are to consider here, is the potential deterrent influence of other individuals' reactions to one's own attempts to manipulate the system. Specifically, Pattanaik notes that the possibility of manipulation of a voting scheme by a single individual (a threat) may be countered by the possibility that other members of society react to it by

0022-0531/80/050267-08$02.00/0

changing their strategies in a harmful way for the threatener (a counterthreat). In order to incorporate considerations of this kind into the analysis, Pattanaik proposes a generic definition of stability, where a voting scheme is stable if its outcomes are such that, for any preference profile, every threat has a counterthreat. This definition allows for a number of specifications, depending upon the restrictions we want to impose on the coalitions who may exert such counterthreats. Gibbard and Satterthwaite's strategy-proofness would be a polar case of stability (strong stability) for which no counterthreatening coalitions are admissible. For a wider notion of stability, where counterthreatening coalitions must be composed of individuals who would not harm themselves by actually carrying out their countherthreats, Pattanaik has shown that there is still no nonimposed, nondictatorial and monotonic voting scheme which is stable [4].

This paper explores the question whether a similar negative result still holds when we further weaken the notion of stability to what we call weak stability, by dropping any restriction on counterthreatening coalitions. As pointed out by Pattanaik, who leaves it as an unresolved problem, "the question is of some interest since (this form of) stability is one of the weakest stability concepts one can think of in this context."[1] But the interest of our notion goes beyond this formal justification. One could think at first sight that counterthreats exerted by coalitions including individuals who would harm themselves by carrying them out would command little credibility, and their deterrent effects be very weak. But there are at least two reasons why such counterthreats may be important. One is that, much in the same way as an Arrowvian "constitution," a voting scheme is to be thought of as a rather permanent arrangement for collective decision-making. And society members may be inclined to "pay the price" of participating in counterthreatening coalitions which go against their immediate interest if this is to help the actual enforcement of the "rules of the game." On the other hand, note that when we consider the possibilities open to an individual as determined by the Gibbard–Satterthwaite result, it is immaterial whether or not the assumed voting intentions of other individuals correspond to their true preferences. Also, information on other individuals will at best be about their actions, not their actual preferences. Thus, an individual who considers the possibility of carrying out a threat need not know whether or not others would harm themselves by countherthreatening, even in the case where he was quite certain about their initial voting intentions.

Our first results are proofs by example that (when individual indifference among alternatives is not allowed) there exist "reasonable" voting schemes which are weakly stable in the sense considered here. The examples we use in proving this fact suggest the appropriateness of still another concept of

[1] The question is also raised and left as an unsolved problem by Kelly [3, Chap. 10].

stability, which we call protective stability. We then show that only voting schemes where some individual has veto power can be stable in this new sense (p-stable). Finally, we discuss the general case (allowing for individual indifferences), and check that the basic results still hold. The paper is organized as follows. Section II furnishes notation, definitions and a restatement of the Gibbard–Sattertywaite and the Pattanaik theorems. Section III contains the examples which prove the existence of "reasonable" weakly stable voting schemes. In Section IV we present the result on protective stability. Section V considers the extension to weak preference orderings.

II. Notation, Definitions and Preceding Results

Let A be a finite set. Elements of A are called the *alternatives* and are denoted by $x, y, z, \ldots$.

Let $I = \{1, 2, \ldots, n\}$ be an initial segment of the integers. Elements of I are called the *individuals*.

Let $\mathscr{R}$ be the set of complete, reflexive and transitive binary relations on A. Elements of $\mathscr{R}$ are called *preference orderings* and are denoted by $R, R', R'', R_i, R_j, \ldots$. The strict preference relation P and the indifference relation I induced by $R \in \mathscr{R}$ are defined in the customary way.

Let $\bar{\mathscr{R}} \subset \mathscr{R}$ be the set of antisymmetric preference orderings. Elements of $\bar{\mathscr{R}}$ are called *linear preference orderings*; they represent preferences under which no two distinct alternatives are indifferent.

Let $\mathscr{R}^n$ be the n-fold cartesian product of $\mathscr{R}$. Elements of $\mathscr{R}^n$ are called *preference profiles* and are denoted by $\mathbf{R}, \mathbf{R}', \ldots$. Elements of $\bar{\mathscr{R}}^n$ are called *linear preference profiles*. When this does not lead to confusion, the ith element of a profile $\mathbf{R}$ will be denoted by R_i, that of a profile $\mathbf{R}'$ by R_i', etc.; also, the strict and indifference relations induced by R_i, R_i', will be denoted by P_i, P_i', I_i, I_i', respectively.

Given $\mathbf{R} \in \mathscr{R}^n$, $R_i' \in \mathscr{R}$, let $\mathbf{R}/R_i'$ denote the preference profile $\mathbf{R}''$, where $R_i'' = R_i'$ and $(\forall j \neq i)\ R_j'' = R_j$.

Given $R \in \mathscr{R}$ and $Y \subseteq A$, *the choice set of R in Y* is the set $C(R, Y) = \{x \mid (\forall z \in Y) xRz\}$. Clearly, when $R \in \bar{\mathscr{R}}$, $Y \neq \varnothing$, $C(R, Y)$ is singleton.

A *general voting scheme* is a function f: $\mathscr{R}^n \to A$.

A *strict voting scheme* is a function f: $\bar{\mathscr{R}}^n \to A$.

The following definitions are stated for general voting scheme. They immediately apply to strict voting schemes by changing $\mathscr{R}$ to $\bar{\mathscr{R}}$, and $\mathscr{R}^n$ to $\bar{\mathscr{R}}^n$.

Let r_f stand for the range of f. f is *dictatorial* iff $\exists i \in I$ such that $(\forall \mathbf{R} \in \mathscr{R}^n)\, f(\mathbf{R}) \in C(R_i, r_f)$.

A voting scheme f has a *vetoer* iff $\exists i \in I$, $\exists x \in A$ such that $(\forall \mathbf{R} \in \mathscr{R}^n)$ $[(\forall z)\ zP_i x \to x \neq f(\mathbf{R})]$. We then say that i has veto power on x under f.

f is *nonimposed iff* $r_f = A$.

f is *monotonic* iff $(\forall x \in A)\{[f(\mathbf{R}) = x \wedge (\forall z, w \neq x)(zR_i w \leftrightarrow zR'_i w) \wedge (\forall z \neq x)(xR_i z \to xR'_i z \wedge xP_i z \to xP'_i z)] \to f(R'_i) = x\}$.

A *threat* to f is a pair $(\mathbf{R}, R'_i) \in \mathscr{R}^n \times \mathscr{R}$ such that $f(\mathbf{R}/R'_i)\ P_i f(\mathbf{R})$. If $(\mathbf{R}, R'_i)$ is a threat to f, we say that i has a threat to f at $\mathbf{R}$.

Let $(\mathbf{R}, R'_i)$ be a threat to f. A *counterthreat* to $(\mathbf{R}, R'_i)$ is a profile $\mathbf{R}'' \in \mathscr{R}^n$ such that $R''_i = R'_i$ and $f(\mathbf{R})\, P_i f(\mathbf{R}'')$. If, in addition, $\mathbf{R}''$ is such that $[f(\mathbf{R}/R'_i)\, R_j f(\mathbf{R}'')] \to [R''_j = R_j]$, we say that $\mathbf{R}''$ is a *strong counterthreat* to $(\mathbf{R}, R'_i)$.

Intuitively, all individuals actively engaged in carrying out a *strong* counterthreat are required to benefit from doing so, with no such restriction being imposed on counterthreats.

A voting scheme f is *strongly stable* if no individual has a *threat* to it at any profile.

A voting scheme f is *stable* if every threat to f has a *strong counterthreat*.

A voting scheme f is *weakly stable* if every threat to f has a *counterthreat*.

Clearly,

$$f \text{ is strongly stable} \to f \text{ is stable} \to f \text{ is weakly stable.}^2$$

The following theorems provide important information on the characteristics of strongly stable and stable voting schemes. They apply to strict voting schemes as well.

THEOREM 1 (Gibbard and Satterthwaite). *Let r_f consist of more than two alternatives. If f is nonimposed and strongly stable, it is dictatorial.*

THEOREM 2 (Pattanaik). *Let r_f consist of more than two alternatives. If f is nonimposed, monotonic and stable, it is dictatorial.*

III. WEAKLY STABLE STRICT VOTING SCHEMES

In view of Theorems 1 and 2, it is natural to inquire whether a further relaxation of our stability requirement to weak stability would still lead to a similar impossibility result. The following examples show that this is not the case for strict voting schemes.

[2] Strong stability is Gibbard and Satterthwaite's strategy proofness. Strong stability, stability and weak stability would be respectively, in Pattanaik's terms, 1-stability (Type 1), 1-stability (Type 3) and 1-stability (Type 2). Pattanaik's terminology is able to encompass a lot of variants of stability which we are not considering here.

EXAMPLE 1. Let A consist of four alternatives and $I = \{1, 2, 3\}$. Consider the strict voting scheme f_1, whose outcome is obtained as follows: eliminate 1's worst alternative from A, then discard the alternative which is worse for 2 among those not eliminated yet, and let the final outcome be the alternative which is best for 3 out of the two which are still left.

We claim that voting scheme f is weakly stable. The reader can check this assertion by noting that individual 3 has no threat to f_1, and that, whenever 1 or 2 have a threat, it involves exerting their veto power against a different alternative than the one they would be excluding by declaring their true preferences. But then, the latter could become the outcome under an appropriate modification of the other individuals' strategies.

A somewhat unsatisfactory feature of this procedure is that all individuals are given a certain veto power, in the sense that anyone can guarantee that an alternative will not be the outcome by just declaring it to be his worst. This will not be the case in our second example.

EXAMPLE 2[3]. Let A consist of three alternatives and $I = \{1, 2, 3, 4, 5\}$. Let T be an arbitrary ranking of the alternatives, to be used as a tie-breaking procedure. Consider the strict voting scheme f_2 defined as follows: each alternative receives one point every time it is not last for one individual, and no points otherwise. The alternative with a larger number of points is the outcome. In case of a tie, first consider whether one of the tied alternatives is Pareto superior to the other (the three alternatives cannot tie). If this is the case, select this Pareto superior alternative to be the outcome. Otherwise, pick among the tied alternatives that which stands higher in ranking T.

Again, f_2 is weakly stable. This is because, on the one hand, no individual has a threat to f_2 at a profile where the outcome is his worst alternative. While, on the other hand, anybody's worst alternative will be the outcome provided each of the remaining alternatives is declared worst by two of the other members of society.

IV. PROTECTIVE STABILITY

Remark that, under both f_1 and f_2, when an individual has a threat and carries it out, he risks ending up with his worst alternative as the outcome, since there exists a counterthreat to any threat. But there is an interesting difference between the two procedures.

Under f_1 every individual guarantees himself by voting sincerely that his worst alternative will not be selected, even if others, by mistake or malice,

[3] This is a member of the class of "approval voting" procedures described by Brams and Fishburn.

carry out on unwarranted action against him. Whereas, under f_2 an individual is always open to the possibility that the actions of other might force the outcome to be his least prefered alternative, whether or not he refrains from carrying out his eventual threats.

In that sense, procedure f_1 provides with stronger incentives than f_2 for individuals to stick to the truth, since it protects sincere individuals from undesirable outcomes which would become possible under preference misrepresentation. We call this further property of f_1 *protective stability* (p-stability) and proceed to formalize it for general voting schemes. (Its definition for strict voting schemes results from the appropriate restriction to admissible profiles).

DEFINITION.[4] A general voting scheme f is p-stable iff every threat $(\mathbf{R}, R_i')$ has a counterthreat $\mathbf{R}''$ such that, for all profiles $\mathbf{R}'''$ $(R_i''' = R_i) \rightarrow f(\mathbf{R}''') P_i f(\mathbf{R}'')$.

This is undoubtedly a strong condition, yet a very attractive one; and we know of at least one strict voting scheme that satisfies it. Theorem 3 below show that only schemes which provide individuals with veto power can be p-stable.

THEOREM 3. *Let f be a p-stable strict voting scheme and let r_f consist of more than two alternatives. Then f has a vetoer. If f is nondictatorial, all individuals whob have a threat to f are vetoers.*

Proof. If f is dictatorial, the dictator is clearly a vetoer. Suppose f is nondictatorial. Then it is not strongly stable, and some individual i has a threat $(\mathbf{R}, R_i')$ to f, in such a way that $f(\mathbf{R}/R_i') P_i f(\mathbf{R})$.

Since f is p-stable by assumption, there must be some counterthreat $\mathbf{R}''$ to $(\mathbf{R}, R_i')$ such that, while $R_i'' = R_i'$ and $f(\mathbf{R}) P_i f(\mathbf{R}'')$ it is also true that, for all profiles $\mathbf{R}''$ under which $R_i'' = R_i$, $f(\mathbf{R}''') P_i f(\mathbf{R}'')$.

Suppose now that i had no veto power over $f(\mathbf{R}'')$, i.e., that for some $\mathbf{R}^{iv}$ we had $f(\mathbf{R}^{iv}) = f(\mathbf{R}'') \equiv w$, while $(\forall x)\, x P_i^{iv} w$. Individual i would then have a threat to f at R^{iv} via R_i. Furthermore, since w is i's worse alternative at R^{iv}, this threat would have no counterthreat, in contradiction to our assumption that f is p-stable.

[4] Under a different interpretation, where individuals would be protecting themselves from uncertainty rather than counterthreats, p-stability would be very similar to requiring that the truth be a unique maximum strategy for all individuals. I am grateful to A. Mas-Colell for calling my attention to this point.

V. Weak Stability and Protective Stability under General Voting Schemes

How do our conclusions change when we consider general rather than strict voting schemes? Since f_1 and f_2 in Examples 1 and 2 are weakly stable strict voting schemes, one may explore the question by extending f_1 and/or f_2 to profiles where individual preferences with indifference among alternatives are allowed. Our definition of f_2 in Example 2 can be applied to the general case by just admitting that more than one alternative can be last for one individual. Clearly, the general voting scheme so defined is also weakly stable. Thus, *our conclusion still holds that Pattanaik's negative result cannot be strengthened to the case of weakly stable schemes.*

However, the reader may check that "natural" extensions of f_1 are no longer weakly stable. Since f_1 is not only weakly stable, but also p-stable, one can think that the difficulty in extending f_1 may be due to the impossibility of carrying through this more stringent condition into the larger framework. A final example shows that this is not the case, by exhibiting a p-stable general voting scheme.

Example 3. Let $A = \{a, b, c, d\}$ and $I = \{1, 2, 3\}$. Consider the voting scheme f_3 defined an follows. Given the individuals' preferences, eliminate all alternatives which are strongly Pareto dominated by some other (all individuals strictly prefer the latter); let A' be the set of nondominated alternatives. If $a \in A'$, and it is not among 1's worst alternatives, let a be the outcome. Otherwise, if $b \in A'$, and it is not among 2's worst alternatives, let b be the outcome. Otherwise, if $c \in A'$ and it is not among 3's worst alternatives, let c be the outcome. Otherwise, let d be the outcome. It is left be the reader to check that f_3 is well defined and p-stable.[5]

Acknowledgments

The author gratefully acknowledges comments from H. Sonnenschein, F. Grafe and an anonymous referee, who pointed out a slip in the proof of Theorem 3 and suggested that the case where individual indifference is allowed should be considered.

[5] Remark that f_3 does no longer treat equally all alternatives (non-neutrality); also, while it always selects alternatives which are not strictly Pareto dominated (thus meeting Arrow's condition P), its outcomes may not be Pareto-optima in the usual sense. For example, where $a\,I_1\,c\,I_1\,d\,P_1\,b$; $b\,P_2\,c\,P_2\,d\,P_2\,a$; $b\,P_3\,c\,P_3\,d\,P_3\,a$, the outcome would be a; yet a is Pareto dominated by c and d. I would conjecture that there is no neutral, Paretian, p-stable, nondictatorial general voting scheme.

REFERENCES

1. R. BRAMS AND P. C. FISHBURN, Approval voting, *Amer. Pol. Sci. Rev.* 72 (September 1978), 831–847.
2. A. GIBBARD, Manipulation of voting schemes: A general result, *Econometrica* **41** (July 1973), 587–601.
3. J. KELLY, "Arrow Impossibility Theorems," Academic Press, New York, 1978.
4. P. PATTANAIK, Counter-threats and strategic manipulation under voting schemes, *Rev. Econ. Studies* **43** (February 1976), 11–18.
5. M. SATTERTHWAITE, Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions, *J. Econ. Theory* **10** (April 1975), 187–217.

Printed by the St. Catherine Press Ltd., Tempelhof 37, Bruges, Belgium