EXCMO. SR. DR. ERIC MASKIN

Discurso de presentación

Dr. Juan Francisco Corona Ramon
Académico de Número
Real Academia Europea de Doctores

Your Excellency, Mr. President
Your Excellencies, Academicians,
Ladies and Gentlemen:

The Royal European Academy of Doctors is pleased to welcome as honorary academician Dr. Eric Stark Maskin, and I am honoured to be asked to give the speech in reply on the solemn occasion of his entry into our much loved corporation, which is proud today to welcome him as honorary academician.

Unfortunately, I did not have the good fortune to receive teaching directly from Professor Maskin, but he has been a presence throughout my academic career, in particular during the years when I was directly involved in the European Public Choice Society, since his studies on Nash Equilibrium and the design of institutions perfectly complemented some of the projects that I had occasion to prepare with my esteemed Professor James Buchanan. Later I also had the opportunity to follow very closely his fundamental contributions to game theory, thanks to the mastery of the much appreciated Andreu Mas-Colell

and, therefore, going on from his acknowledged academic distinction, it is an emotional moment for me to have the immense honour of replying to his admission speech.

A brief biographical sketch

Eric S. Maskin was born in New York City although he grew up further north, in Alpine, a little town beside the Hudson River. He went to secondary school at Tenafly, three miles from home, where thanks to his calculus teacher he discovered the beauty of mathematics. So much so that Maskin decided to study mathematics at Harvard University, where he became one of their finest alumni. There he shared algebra classes with Pierre Samuel and Richard Brauer and analysis with George Mackey and Lars Ahlfors, in whom he found great inspiration.

His first contact with economics was almost accidental, when he attended a course on "economics of information" given by Kenneth Arrow, who later would be his doctorate tutor. In this course, Maskin discovered the work of Leonid Hurwicz on the incipient field of mechanism design. This work was a revelation to him. In his own words, "it had the precision, the rigour and, at times, the beauty of pure mathematics and it was also orientated to problems of real social importance; an irresistible combination".

As a result of this discovery, Maskin did his doctorate in applied mathematics, for which he attended various classes on economics including a course on general equilibrium given by Truman Bewley, where he met his class companion and later co-Nobel prize-winner, Roger Myerson, and also a seminar on analysis by Jerry Green, where he met students of the stature of Elhanan Helpman, Bob Cooter, and Jean-Jacques Laffont.

While doing his doctorate, Maskin learned from his tutor, Ken Arrow. When he finished it, he secured a contract for post-doctoral research with Frank Hahn at Cambridge University. While he was in England, Maskin submerged himself in resolving a new problem: in what circumstances is it possible to design a mechanism which implements a given social objective. After working on the question for nearly the whole year he came to the conclusion that the key was to be found in monotonicity (a concept that we shall explain later). This discovery was a revelation although the formula was pretty complex. It was then that his friend (disputant) Karl Vind suggested a simplifica-

tion. Maskin then wrote all the details in his article "Nash Equilibrium and Welfare Optimality" when already a professor at MIT, although he did not publish it until twenty years later, since by then it was already known in the format of a working paper.

During his stay at MIT as professor he had the opportunity to come across distinguished personalities such as Paul Samuelson, Franco Modigliani and Bob Solow. At this time he said that he learned much from the teachings of Peter Diamond, who acted as his big brother.

Seven years later Maskin left MIT to take up a place as a professor at Harvard, where he was part of a group of theoreticians among whom were Andreu Mas-Colell, Jerry Green, Oliver Hart, Drew Fudenberg, Mike Whinston, Marty Weitzman. After 15 years in this prestigious but also very demanding university, Eric went to work in the Institute for Advanced Study where he continued his research work as well as giving classes at Princeton University, where until 2011 he held the "Albert O. Hirschman" Chair, and supervised doctorate students.

Up till now, Maskin has written more than 130 books and articles, he is a member of numerous organisations all over the world, among which are: the Econometric Society where he was president in 2003, the European Economic Association and the American Academy of Arts and Sciences, considered as the oldest and most prestigious honorary society and a leading centre of policy research in the United States. Maskin has Honorary Doctorates in various universities of many different countries and has been awarded 15 prizes and top level medals in recognition of his work.

Eric S. Maskin won the Nobel Prize for Economics in 2007 together with Leonid Hurwicz and Roger Myerson "for having laid the foundations of mechanism design theory", studying the design of social decision procedures in situations in which economic agents have private information and use it in a strategic way.

Maskin thought himself lucky to have been able to discover economics, to have entered the field at a moment when mechanism design was beginning to flourish, to have had a series of outstanding maestros, students, colleagues and friends during his career and, most important of all, to be able to devote himself to the profession which he loved.

Given the variety of his contributions, I wanted to divide my reply into two parts: first I will give a brief summary of some of his most significant achievements. Then I shall illustrate the importance of his research with examples of practical problems to which we can apply his results.

Brief summary of E. Maskin's theoretical achievements

I would like to talk about six specific achievements by Eric Maskin with great impact in different fields: (i) fundamental aspects of game theory, (ii) mechanism design for the implementation of social choice, (iii) banking economics, (iv) auction theory, (v) inequality theory and (vi) voting systems.

(i) Fundamental contributions to game theory[1]

In game theory, the folk theorem says that any feasible profile of payments which dominates minimax strategy -consisting of penalising the rival- can be achieved as Nash equilibrium in infinitely repeated games if the time discount factor is sufficiently small.

For example, in the prisoner's dilemma, the only Nash equilibrium occurs when both players betray, which is also a mutual minimax profile. In this case, cooperation is not Nash equilibrium. The 'folk' theorem says that if the players are sufficiently patient, in the infinitely repeated version of the game there is Nash equilibrium of such a manner that both players cooperate.

Maskin and Fudenberg demonstrated that this theorem is also valid in finite repeated games *and with imperfect information* (and mixed strategies). It is only necessary to fulfil one of these two conditions: that there are two players in the game, or that players who do penalties can be rewarded ("full dimensionality condition").

This result is important because until its publication, cooperation actions could only be explained in the framework of a theory of infinitely repeated games; not in a framework of finite games, which contrasts clearly with the fact that economic agents generally have a limited life and that cooperation is often observed in experiments with finite repetitions.

1. The Folk Theorem in Repeated Games with Discounting or with Incomplete Information, published jointly with Drew Fundenberg in Econometrica in 1986.

The strategy which sustains these possible cooperative equilibriums is characterised by the fact that, after a deviation, each player changes to a minimax strategy, penalising the other player for a specific number of periods. With multiple players, those who deviate must be penalised, but also those who penalise must themselves be threatened with sanctions in the case of not penalising those who deviate.

In the cases of imperfect information, a player can credibly threaten with taking suboptimal decisions if there is a (small) probability that the action is in fact optimum because there is interest in maintaining a reputation for possible "irrationality".

In this way, games with a finite and infinite horizon can reproduce the same results. But Maskin's article gives additional interest to games with a finite horizon, because one can argue in favour of or against certain equilibriums depending on the type of irrationality necessary to sustain them.

(ii) Mechanism design for the implementation of social choice[2]

As we have seen, Maskin extended game theory as a conceptual framework in order to explain a great many situations in their finite interaction variants. In consequence, it becomes a basic tool for mechanism design and the implementation of social choice.

In this specific terrain, Maskin identified the conditions necessary for a social choice to be able to be implemented by means of a mechanism compatible with the incentives of the participants.

As I explained in the introduction, he discovered that the key property of social preferences for the implementation of a social decision in the form of Nash equilibrium is "monotonicity". Any rule of social choice which fulfils this condition and the non-existence of the power of veto (by any of the participants) can be implemented by a game or "mechanism" if there are three or more individuals.

The "monotonicity" of social preferences requires that if a result/option is optimum in a given state, it will also be so in any other state in which this option does not lose position in the ranking of preferences of any of the participants

2. **Nash Equilibrium and welfare optimality**, *published in the Review of Economic Studies in 1999.*

in relation to the other options (the order of the rest of the results being able to vary).

(iii) Banking economy: efficiency in the grant of credit in centralised v. decentralised systems[3]

Maskin also applied his knowledge of game theory to explain how financial decentralisation (the number of banks) affects the grant of credit in a framework of adverse selection, in which unprofitable projects are refinanced after having incurred sunk costs (which happen when information on the quality of the project is incomplete and which, had they been foreseen, would have made the project undesirable).

He showed how, when there are multiple banks, they can publicly commit not to refinance unprofitable projects thus discouraging entrepreneurs from seeking finance. This extra in financial discipline is an argument in favour of financial decentralisation. However, in his model, a large number of banks implies an excessive emphasis on profitability in the short term.

Maskin's model is important in explaining the existence of "soft" budgetary restrictions in state financing, which means that state companies easily attain state financing in spite of embarking on bad projects. It can also explain the differences between banking and industrial relationships among Anglo-Saxon types of economies, characterised by high banking dispersion, and those of Germany and Japan, where the financial concentration is greater.

(iv) Auction theory[4]

Maskin explained that when buyers are averse to risk, an auction in which the winner pays the highest price generates more income for the seller than an auction in which the winner pays the second price. Also, the seller's preference for the first type of auction is greater if he also is averse to risk.

This is in contrast with the more normal result, as he himself has pointed out, which indicates that it is optimum to award the lot to the winning buyer at the second highest price.

---

3. **Credit and Efficiency in Centralized and Decentralized Economies,** *jointly with M. Dewatripont published by "Review of Economic Studies", in 1995*
4. **Optimal Auctions with Risk Averse Buyers**, jointly with John Riley, published in Econometrica in 1984

(v) Inequality theory[5]

I would also like to mention one of his more recent contributions in the social field: his inequality theory.

In a few words, Maskin argues that workers qualified in developing countries are sought-after by the multinationals and get salary rises. In contrast, unqualified workers are ignored, so that their salaries tend to fall with globalization.

The message of this theory is not that we need to distance ourselves from globalization to avoid increased inequality. Even if that were possible, to do it would involve a high cost in the long term in terms of income per inhabitant. Rather, Maskin insists that the best remedy for inequality is to give poorly qualified workers the opportunity (and the tools) to share the benefits of globalization.

(vi) Voting systems[6]

According to Gibbard-Satterthwaite's impossibility theorem there is no voting regulation which respects at the same time all the optimum criteria of representativeness, that is: (1) that it cannot be manipulated -that is, that it prevents strategic behaviour by voters to avoid a given candidate; (2) that it maintains the consensus -that is, that if all the voters prefer one candidate A to another B, that B is not elected; (3) that all the votes have the same value; (4) that equal treatment is given to all candidates; (5) that the election always gives a clear result.

But Eric Maskin has argued that this impossibility theorem is too pessimistic. What matters is not whether a voting system can transmit any preference adequately. What matters is whether it transmits a reasonably broad set of preferences and the most likely to occur in reality.

Maskin and Dasgupta have demonstrated that there are 2 regulations for voting which maximise the situations in which a system is "optimum". They

5. Globalization and Inequality, an outline of the article in The Economist (2014)
6. Relevant sources: On the robustness of majority rule (2008), with Dasgupta; Pandering and Pork Barrel Politics (2014), with Jean Tirole; How should we elect our leaders? (2014); Elections and strategic voting (2011).

are the Condorcet Method[7] and the Borda Count[8]. Happily, these two methods of voting are complementary, so that when one fails the other works.

Normally, the Condorcet method satisfies all the criteria of representativeness, but sometimes it is unable to ensure its decision-making capacity (criterion 5), as it can be a victim of what is called "Condorcet cyclical dependency". With 3 candidates, for example, a majority may prefer the some other candidate, for each candidate, if the preferences are sufficiently complicated.

In these cases, however, the Borda Count works properly. In this way, with just two complementary methods an election can be achieved which meets the established requirements.

Practical implications

Mechanism design theory is a fundamental field for the social sciences. It allows solutions to be found for complex problems of collective choice. It is, also, a field with enormous development potential and one which offers practical recommendations.

To have Eric Maskin among us today is an opportunity to thank him for his contributions and to record the enormous advances that he has made possible:

- Development of game theory in the broad sense.
- Inspiring auction systems and the grant of licences throughout the world.
- Notably improving institutional quality and economic efficiency.

But there is still much to do. Specific problems on which there is still much terrain to be explored.

In fact, in Spain, the auctions held of energy and mobile telephony have been based on recommendations derived from Professor Maskin's work. At a practical level, the latest auction of 4G in Spain was assigned with lots at highest price and with maximum limits on what each operator could buy. It can

7. According to which the voters put the candidates in order from greater to lesser preference and the candidate is found who wins all the pairings.
8. According to this methodology, the candidates are ordered according to the preferences of each elector; in the count, points are given for each position in the order: 1 point for last in line, 2 points for the next to last, 3 for the antepenultimate, etc.

be asked whether this is the best system (according to whether we consider the participants as averse to risk or not), but it appears to be an efficient system. On the other hand, Television spaces are assigned in a much less transparent way. Also it is the government that decides what percentage of frequencies are allotted to Mobiles and which to TV, and later they are assigned within each sector and it is impossible to move them from TV to mobiles or vice versa.

Another example of a practical ambit with potential for improvement is the market in greenhouse gas emissions. Before the crisis, rights were awarded to various geographical zones. The EU got a fixed number which should have allowed the creation of a market in rights which would be an incentive to green investments to reduce emissions. However, electrical demand slumped with the crisis and the number of rights distributed was too high. In consequence, there were no incentives to make the necessary investments. After this experience, perhaps we could conclude that a system should be designed in which the number of rights depends dynamically on the GDP or some other variable.

We can also extract implications from Eric Maskin's work for our financial sector. The process of concentration that Spain has been through will have numerous effects: surely the emphasis on grants of credit will lessen in the short term, but, *ceteris paribus*, the tendency to maintain commitments to unprofitable projects could have increased.

Also inequality theory is of great interest in tackling, understanding and motivating the educational reforms necessary for our country to prosper. This theory explains to us why we need to develop and attract talent.

Finally, Eric Maskin has made specific contributions in the study of electoral systems and on the strategic behaviour of voters.

It is a privilege to have Eric among us today, precisely at a time when the Spanish electoral system has become extremely complex. Now the electors will have to vote in a more strategic way, penalising those who obstructed the formation of a government or voting against their less favoured option (and not according to their preferences).

His work teaches us that this strategic behaviour not only distorts the transmission of individual preferences, and therefore, the essence of the democratic system, but also imposes a significant cost on the voter: it is already hard enough to discover which party best represents his preferences; now also we

need to understand game theory and know how to predict the preferences of the other electors in order to vote in consequence. The citizen's decision problem is now much more complex.

In the name of the President of the Royal European Academy of Doctors, in those of all the academicians and in my own, please receive, my dear Eric, our most cordial welcome.

Congratulations.

□ □ □

An Introduction to Mechanism

Dr. Eric Maskin
Harvard University

Theory of Mechanism Design –
"engineering" part of economic theory

- much of economic theory devoted to:

  – understanding existing economic institutions
  – explaining/predicting outcomes that institutions generate
  – positive, predictive

- mechanism design – reverses the direction

– begins by identifying desired outcomes (goals)
– asks whether institutions (mechanisms) could be designed to achieve goals
– if so, what forms would institutions take?
– normative, prescriptive

For example, suppose

- mother wants to divide cake between 2 children, Alice and Bob
- goal: divide so that each child is happy with his/her portion
  - Bob thinks he has got at least half
  - Alice thinks she has got at least half call this fair division
- If mother knows that the kids see the cake in same way she does, simple solution:
  - she divides equally (in her view)
  - gives each kid a portion
- But what if, say, Bob sees cake differently from mother?
  - she thinks she's divided it equally
  - but he thinks piece he's received is smaller than Alice's
- difficulty: mother wants to achieve fair division
  - but doesn't have enough information to do this on her own
  - in effect, doesn't know which division is fair
- Can she design a mechanism (procedure) for which outcome will be a fair division?

  (even though she doesn't know what is fair herself ?)
- Age-old problem
  - Lot and Abraham dividing grazing land

Age-old solution:

- have Bob divide the cake in two
- have Alice choose one of the pieces

Why does this work?

- Bob will divide so that pieces are equal in his eyes
  - if one of the pieces were bigger, then Alice would take that one
- So whichever piece Alice takes, Bob will be happy with other
- And Alice will be happy with her own choice because if she thinks pieces unequal, can take bigger one

Example illustrates key features of mechanism design:

- mechanism designer herself doesn't know in advance what outcomes are optimal
- so must proceed indirectly through a mechanism
  - have participants themselves generate information needed to identify optimal outcome
- complication: participants don't care about mechanism designer's goals
  - have their own objectives
- so mechanism must be incentive compatible
  - must reconcile social and individual goals

Second Example:

Suppose government wants to sell right (license) to transmit on band of radio frequencies (real-life issue for many governments, including in U.S.)

- several telecommunication companies interested in license
- goal of government: to put transmitting license in hands of company that values it most ("efficient" outcome)

- but government doesn't know how much each company values it (so doesn't know best outcome)

Government could ask each company how much it values license

- but if company thinks its chances of getting license go up when it states higher value, has incentive to *exaggerate* value

- so no guarantee of identifying company that values it most

- government could have

  − each company make a bid for license
  − high bidder wins license
  − winner pays bid

- but this mechanism won't work either

  − companies have incentive to understate

- suppose license worth $10m to Telemax, then

  – if Telemax bids $10m and wins, gets
  $10m − $10m = 0

- so Telemax will bid less than $10m

- but if all bidders are understating, no guarantee that winner will be company that values license most

Solution:

- every company makes bid for license

- winner is high bidder

- winner pays *second-highest* bid

  − so if 3 bidders and bids are
  $10m, $8m, and $5m,
  winner is company that bids $10m

− but pays only $8m

• Now company has no incentive to understate

− doesn't pay bid anyway
− if understates, may lose license

• Has no incentive to overstate

− If bids $12m, will now win if other company bids $11m
− But overpays

• So best to bid *exactly* what license worth

• And winner will be company that values license most

• Have looked at 2 applications of mechanism design theory

• Many other potential applications

1) International treaty on greenhouse gas emissions

2) Policies to prevent financial crises

3) Design of presidential elections

□ □ □

Trabajos aportados por el nuevo Académico de Honor

Nash Equilibrium and Welfare Optimality

ERIC MASKIN
Harvard University

If A is a set of social alternatives, a social choice rule (SCR) assigns a subset of A to each potential profile of individuals' preferences over A, where the subset is interpreted as the set of "welfare optima". A game form (or "mechanism") implements the social choice rule if, for any potential profile of preferences, (i) any welfare optimum can arise as a Nash equilibrium of the game form (implying, in particular, that a Nash equilibrium exists) and, (ii) *all* Nash equilibria are welfare optimal. The main result of this paper establishes that any SCR that satisfies two properties—monotonicity and no veto power—can be implemented by a game form if there are three or more individuals. The proof is constructive.

I. INTRODUCTION

After society has decided on a social choice rule—a recipe for choosing the optimal social alternative (or alternatives) on the basis of individuals' preferences over the set of all social alternatives—the social planner still faces the problem of how to implement that rule. In particular, the planner may not know individuals' preferences. He might attempt to elicit them, but this may not be an easy task, even abstracting from communication costs. If individuals know the rule by which the planner selects alternatives on the basis of reported preferences, they may have an incentive to report falsely.

One can think of the individuals as playing a game form. They are endowed with strategy spaces coinciding with their sets of possible announcements. The strategies that players choose determine an outcome. Ideally, one might hope to devise game forms which ensure that individuals will always want to announce their true preferences and that the right outcome (*i.e.* the one prescribed by the social choice rule) relative to those preferences is selected. In the case where preferences can be *anything*—that is, when the planner can place no *a priori* restrictions on the nature of individuals' preferences—Gibbard (1973) and Satterthwaite (1975) dash this hope by demonstrating that only dictatorial game forms have the property that players always wish to announce the truth regardless of the strategies of others. In other words, only a game form in which there exists a player who always gets his favourite alternative is strategy-proof.

In view of this negative result, one may be willing to sacrifice the strong incentive-compatibility of strategy-proofness. One may require, for example, only that players be in Nash equilibrium. This weaker stipulation has in fact been pursued by Groves and Ledyard (1977), Hurwicz (1979), and Schmeidler (1980), who construct game forms for the allocation of economic resources—with no restriction on preferences other than the usual convexity, continuity, and monotonicity assumptions—such that Nash equilibria exist and are Pareto optimal. Moreover the game forms constructed are nondictatorial. Indeed, in the Hurwicz and Schmeidler papers the Nash equilibria are not only Pareto optimal, but coincide with the set of Walrasian or Lindahl equilibria.

In this paper I examine the general question of implementation of social choice rules by game forms when Nash equilibrium is the solution concept. The main result asserts that a social choice rule on an arbitrary domain of preferences can be implemented by a game form if it satisfies two arguably reasonable properties: *monotonicity* and *no veto power*.

As I have presented it, implementation theory may appear to be purely a topic in applied welfare economies: Given the desired SCR, how can we go about implementing it? But there is a *positive* aspect to the theory as well. Certain well-known mechanisms—*e.g.* the English auction in the context of selling goods or rank-order voting in the context of electing candidates—are used frequently in practice, and we may wonder what properties the outcomes they give rise to satisfy as individuals' preferences vary. This is a question that the theory can also answer.

I proceed as follows. In the second section I introduce most of the notation and definitions. In the third, I present an "impossibility" result for the case of two players. In the fourth, I discuss the properties of monotonicity and no veto power. I demonstrate that monotonicity is an essential requirement of a social choice rule for implementability. I suggest also that no veto power, though not a necessary condition, is really quite weak and, in fact, is vacuously satisfied in many contexts.

Then in Section V, I present the main result of the paper: a constructive demonstration that monotonicity and no veto power suffice for an SCR of more than two individuals to be implementable in Nash equilibrium. I also show, by example, that the result does not remain true if we drop the no veto power hypothesis.

Finally, in Section VI, I show that we can retain implementability with an even weaker no veto power condition if we impose an individual rationality requirement on the SCR.

II. DEFINITIONS AND NOTATION

Let A be a non-empty, possibly infinite set of social alternatives and let $\mathscr{R}_A$ be the class of all orderings of the elements of A ($\mathscr{R}_A$ is sometimes called the *unrestricted domain* of preferences). If $\mathscr{R}_1, \ldots, \mathscr{R}_n$ are sub-classes of $\mathscr{R}_A$, where n is a positive integer, then f is an n-person *social choice rule* (SCR) on $\mathscr{R} = \mathscr{R}_1 \times \cdots \times \mathscr{R}_n$ if f is a correspondence

$$f: \mathscr{R} \twoheadrightarrow A.$$

One interprets an SCR as selecting a set of "welfare optimal" alternatives $f(\boldsymbol{R})$ for each profile of preferences $\boldsymbol{R} = (R_1, \ldots, R_n) \in \mathscr{R}$, where $R_i (\in \mathscr{R}_i)$ is individual i's preference ordering of A, and $\mathscr{R}_i$ is his domain of possible preference orderings. If $a \in f(\boldsymbol{R})$, we say that a is *f-optimal* for profile $\boldsymbol{R}$.

Prominent examples of SCRs include (i) the (weak) *Pareto correspondence*, which selects all weak Pareto optima corresponding to given profile $\boldsymbol{R}$:

$$f^{\mathrm{PO}}(\boldsymbol{R}) = \{a | \text{for all } b \in A \text{ there exists } i \text{ such that } aR_ib\};^1$$

(ii) the *Condorcet correspondence*, which, for each profile $\boldsymbol{R}$ of strict preferences,[2] selects each alternative that a (weak) majority prefers to any other alternative:

$$f^{\mathrm{CON}}(\boldsymbol{R}) = \{a | \text{for all } b \in A \ \#\{i | aR_ib\} \geq \#\{i | bR_ia\}\};^3$$

1. The notation "aR_ib" means "a is at least as high as b in the ordering R_i" (*i.e.* a is weakly preferred to b).

2. A preference ordering is *strict* if it ranks no two alternatives as indifferent.

3. The notation $\#\{i | aR_ib\}$ denotes the number of individuals who prefer a to b (recall that we are dealing with strict preferences).

and, in a pure exchange economy of l goods, where an alternative a constitutes an *allocation* of goods across individuals (*i.e.* $a = (a_1, \ldots, a_n)$, where $a_i \in \mathbb{R}^l_+$), (iii) the *Walrasian correspondence*, which, given individuals' endowments $(\omega_1, \ldots, \omega_n)$, chooses the set of allocations that can arise in competitive equilibrium:

$$f^W(R) = \{a \mid \Sigma a_i = \Sigma \omega_i \text{ and there exists a price vector } p \in \mathbb{R}^l_+ \text{ such that,}$$
$$\text{for all } i,\ a_i \in \mathbb{R}^l_+, p \cdot (a_i - \omega_i) = 0 \text{ and if}$$
$$\text{for some } b_i \in \mathbb{R}^l_+,\ b_i P(R_i) a_i[4]$$
$$\text{then } p \cdot (b_i - \omega_i) > 0\}.$$

An SCR differs from a *social welfare function* in the sense of Arrow (1951) (a mapping $F: \mathscr{R} \to \mathscr{R}_A$) in that it does not rank non-optimal alternatives. Clearly, however, a social welfare function F induces a natural social choice rule: the correspondence which selects the alternatives top-ranked by F for each profile.

Given strategy spaces $S_1, \ldots, S_n$, an *n-person game form* ("mechanism" and "outcome function" are two synonyms) g *on* A is a mapping

$$g: S_1 \times \cdots \times S_n \to A.$$

If players 1 through n choose strategies s_1 through s_n, respectively, then alternative $g(s)$, where $s = (s_1, \ldots, s_n)$, is the outcome. Moreover, if players use the vector of mixed strategies $\mu = (\mu_1, \ldots, \mu_n)$[5] we denote the random outcome by $g(\mu)$, for which the probability of outcome $g(s_1, \ldots, s_n)$ is $\mu_1(s_1) \cdots \mu_n(s_n)$.

We say that the game form g *implements* the social choice rule f *in Nash equilibrium* if and only if

$$\forall R = (R_1, \ldots, R_n) \in \mathscr{R}\ \forall a \in f(R) \text{ there exists}$$
$$s = (s_1, \ldots, s_n) \in \prod_{i=1}^{n} S_i \text{ such that } g(s) = a \text{ and}$$
$$g(s) R_i g(s'_i, s_{-i})[6] \text{ for all } i \in \{1, \ldots, n\} \text{ and all } s'_i \in S_i; \tag{1}$$

and

$$\forall R \in \mathscr{R} \text{ if } \mu \text{ is a mixed-strategy Nash equilibrium}[7] \text{ of } g \text{ with respect to}$$
$$R \text{ then } g(s) \in f(R) \text{ for all realizations } s \text{ in the support of } \mu. \tag{2}$$

Requirement (1) needs little explanation if one's solution concept is Nash equilibrium. It states simply that any welfare-optimal alternative (as defined by f) can arise as a (pure-strategy) Nash equilibrium of the game form.[8] We could alternatively impose the weaker requirement that, for all $R \in \mathscr{R}$, there exists *some* $a \in f(R)$ for which there is a Nash equilibrium of g resulting in a. But this would not lead to significantly different results. Indeed, if the game form g implements f using the alternative condition in place of (1),

4. $xP(R_i)y$ means that x is strictly preferred to y under R_i.
5. A mixed strategy μ_i for player i assigns a probability $\mu_i(s_i)$ to each (pure) strategy s_i.
6. The notation "$g(s'_i, s_{-i})$" denotes $g(s_1, \ldots, s_{i-1}, s'_i, s_{i+1}, \ldots, s_n)$.
7. If μ and μ' are nondegenerate mixed strategy vectors, then player i's preference between $g(\mu)$ and $g(\mu')$ may not be fully specified by his ordinal ranking R_i; we may have to know his risk preferences as well. However, the analysis in this paper holds for any risk preferences consistent with R_i.
8. Requirement (1) is essentially the stipulation that the game form be *unbiased* (see Hurwicz (1979*b*)).

define the subcorrespondence f' as $f'(\boldsymbol{R}) = \{a \in f(\boldsymbol{R}) |$ there exists an equilibrium of g resulting in $a\}$. Then g implements f' in the standard sense (*i.e.* using (1)).

Requirement (2), which is essentially the converse of (1), is also quite natural. Given that, in general, there can be multiple Nash equilibria of g and that, in the absence of a theory of how players select among these, one cannot predict which of these will ultimately arise, requirement (2) is necessary to ensure f-optimality of the outcome.

III. THE TWO PLAYER CASE

One might well argue that most social choice rules of interest satisfy the Pareto property.

Pareto property. The SCR f: $\mathscr{R} \to\!\!\!\!\to A$ satisfies the Pareto property if, for all $\boldsymbol{R} \in \mathscr{R}$, $f(\boldsymbol{R}) \subseteq f^{\mathrm{PO}}(\boldsymbol{R})$.

We shall see that the prospects for implementing two-person Pareto optimal SCRs on an unrestricted domain of preferences are quite bleak. We need the following definition:

Dictator. An individual i is a dictator for an SCR

f: $\mathscr{R} \to\!\!\!\!\to A$ if and only if

$$[\forall \boldsymbol{R} \in \mathscr{R} \forall a \in A,\ a \in f(\boldsymbol{R}) \text{ if and only if } aR_i b \text{ for all } b \in A].$$

In other words, individual i is a dictator if, for any profile of preferences, the set of welfare-optimal alternatives (with respect to f) consists of the *top-ranked* alternatives for i (the alternatives that i prefers to any other). An SCR that has a dictator shall be called *dictatorial.*

I now show that any Pareto-optimal two-person SCR that is implementable must be dictatorial if it is defined on the unrestricted domain of preferences.

Theorem 1. *Let f: $\mathscr{R}_A \times \mathscr{R}_A \to\!\!\!\!\to A$ be a two-person* SCR *satisfying the Pareto property. Then f can be implemented if and only if it is dictatorial.*[9,10]

Proof. First observe that if f is dictatorial, it is trivially implementable; if i is the dictator, just take the game form in which player i announces an alternative and his announcement is implemented.

To prove the proposition in the other direction, suppose that $g: S_1 \times S_2 \to A$ implements f. If A contains only one element, the result is trivial. Therefore assume that A contains at least two elements. For each $s_2^* \in S_2$, let $T_1(s_2^*) = \{a \in A | g(s_1, s_2^*) \neq a$, for all $s_1 \in S_1\}$. Define $T_2(s_1^*)$ for $s_1^* \in S_1$ analogously. Notice that $T_i(s_j^*)$ is the set of alternatives that player i cannot induce, given that player j's $(j \neq i)$ strategy is s_j^*.

Claim 1. *For any $s_1 \in S_1$ and $s_2 \in S_2$, $T_1(s_2) \cap T_2(s_1) = \phi$. That is, starting from any pair of strategies (s_1, s_2), any alternative a can be reached by a unilateral deviation by some player.*

9. This result has also been obtained, in somewhat different form, by Hurwicz and Schmeidler (1978).

10. Theorem 1 remains true if we replace $\mathscr{R}_A$ with the somewhat smaller domain $\mathscr{R}_A^*$ of *strict* orderings.

Proof of Claim 1. Suppose for $\bar{s}_1 \in S_1$ and $\bar{s}_2 \in S_2$, there exists $a \in T_1(\bar{s}_2) \cap T_2(\bar{s}_1)$. Take $b = g(\bar{s}_1, \bar{s}_2)$. Then, $b \neq a$ by construction. Choose $(\bar{R}_1, \bar{R}_2) \in \mathscr{R}_A \times \mathscr{R}_A$ such that, for all $i \in \{1, 2\}$ and all $c \in A \backslash \{a, b\}$, $aP(\bar{R}_i)bP(\bar{R}_i)c$. Observe that $(\bar{s}_1, \bar{s}_2)$ constitutes a Nash equilibrium for preferences $(\bar{R}_1, \bar{R}_2)$, yet b is not Pareto optimal, a contradiction of f's Pareto optimality. Hence $T_1(\bar{s}_2) \cap T_2(\bar{s}_1) = \phi$. ‖

Claim 2. *For all $a \in A$, if $a \notin \bigcup_{s_2 \in S_2} T_1(s_2)$, then there exists $\hat{s}_1 \in S_1$ such that, for all $s_2 \in S_2$, $g(\hat{s}_1, s_2) = a$. Similarly, if $a \notin \bigcup_{s_1 \in S_1} T_2(s_1)$, then there exists $\hat{s}_2 \in S_2$ such that $\forall s_1 \in S_1$, $g(s_1, \hat{s}_2) = a$. That is, if no strategy by player 2 prevents player 1 from inducing alternative a, then player 1 has a strategy that guarantees alternative a (and similarly for player 2).*

Proof of Claim 2. It suffices to prove the statement about player 1's strategy $\hat{s}_1$. Consider $a \in A$ such that $a \notin \bigcup_{s_2 \in S_2} T_1(s_2)$. Choose $\hat{R}_1, \hat{R}_2 \in \mathscr{R}_A$ such that $\forall b \in A \backslash \{a\}$, $aP(\hat{R}_1)b$ and $bP(\hat{R}_2)a$. Let $(\hat{s}_1, \hat{s}_2)$ be a Nash equilibrium for $(\hat{R}_1, \hat{R}_2)$. Because $a \notin T_1(\hat{s}_2)$, there exists $s_1 \in S_1$ such that $g(s_1, \hat{s}_2) = a$. For $(\hat{s}_1, \hat{s}_2)$ to be a Nash equilibrium, therefore, we must have $g(\hat{s}_1, \hat{s}_2) = a$. Suppose there exist $b \in A \backslash \{a\}$ and $s_2 \in S_2$ such that $g(\hat{s}_1, s_2) = b$. Then from our choice of $\hat{R}_2$, $(\hat{s}_1, \hat{s}_2)$ cannot be a Nash equilibrium. We infer that, $\forall s_2 \in S_2$, $g(\hat{s}_1, s_2) = a$, as desired. ‖

Now, for any $a \in A$, either $a \notin \bigcup_{s_2 \in S_2} T_1(s_2)$ or $a \notin \bigcup_{s_1 \in S_1} T_2(s_1)$, otherwise Claim 1 is violated. Suppose there exist $a, b \in A$, $a \neq b$, such that $a \notin \bigcup_{s_2 \in S_2} T_1(s_2)$ and $b \notin \bigcup_{s_1 \in S_1} T_2(s_1)$. By Claim 2, there exist $\bar{s}_1 \in S_1$ and $\bar{s}_2 \in S_2$ such that $\forall s_2 \in S_2$, $g(\bar{s}_1, s_2) = a$ and $\forall s_1 \in S_1$, $g(s_1, \bar{s}_2) = b$. But then $g(\bar{s}_1, \bar{s}_2) = a$ and $g(\bar{s}_1, \bar{s}_2) = b$, which is impossible. Therefore, either $\forall a \in A$, $a \in \bigcup_{s_2 \in S_2} T_1(s_2)$ or $\forall a \in A$, $a \notin \bigcup_{s_1 \in S_1} T_2(s_1)$. The first statement implies, by Claim 2, that 1 is a dictator for f, the second that player 2 is a dictator. ‖

The negative conclusion of Theorem 1 depends on there being an unrestricted domain of preferences. For restricted domains, results can be quite positive, *e.g.* in the case of "economic preferences," where preferences are required to be increasing, continuous, and convex over allocations of a divisible good (see Dutta and Sen (1991) and Moore and Repullo (1990) for a complete characterization of the implementation possibilities in the $n = 2$ case).

III(i). *An example with more than two players*

When $n > 2$, the conclusion of Theorem 1 no longer holds. Indeed, for this case, it is possible to implement Pareto optimal and nondictatorial SCRs defined on the unrestricted domain of preferences. Consider the following example.

Example 1.[11] For any positive integers m and n, take $A = \{a_1, \ldots, a_m\}$, $S_1 = \{2, \ldots, n\}$, and $S_2 = \cdots = S_n = A$. Define g: $S_1 \times \cdots \times S_n \to A$ so that $\forall (s_1, \ldots, s_n) \in \prod_{j=1}^{n} S_j$, $g(s_1, \ldots, s_n) = s_{s_1}$. That is, player 1 chooses a player s_1 (other than himself), and player s_1 chooses the outcome from A. I claim that this game form implements the SCR $f^{\mathrm{KM}}(R) = \{a |$ there exists $j \in \{2, \ldots, m\}$ such that aR_jb for all $b \in A\}$ when players have preferences in $\mathscr{R}_A$. In other words, f^{KM} chooses each alternative for which there exists some individual other than 1 who top-ranks it.

11. This example is adapted from Hurwicz and Schmeidler (1978), who call the game form we have constructed the "king-maker" mechanism.

To see that g implements f^{KM}, we first note that, for any profile $\boldsymbol{R} = (R_1, \ldots, R_n)$, each $a \in f^{KM}(\boldsymbol{R})$ corresponds to some Nash equilibrium of g. In particular, if $a \in f^{KM}(\boldsymbol{R})$ there exists some $j \in \{2, \ldots, m\}$ such that a is top-ranked by R_j. Then the strategy profile $(j, a, \ldots, a)$ is a Nash equilibrium for $\boldsymbol{R}$, and $g(j, a, \ldots, a) = a$, as required. Hence, it remains only to show that all Nash equilibria of g are f-optimal. Suppose $(\mu_1, \ldots, \mu_n)$ is a mixed-strategy equilibrium of g with respect to $\boldsymbol{R}$. Consider $j \in \{2, \ldots, n\}$ to which μ_1 assigns positive probability. Then player j has a positive chance of being able to choose the alternative he wants. Therefore for μ_j to be an equilibrium strategy, it must assign positive probability only to player j's top-ranked alternatives. But this means that only an alternative that is top-ranked for some individual among $2, 3, \ldots, n$ can be a realization of $g(\mu_1, \ldots, \mu_n)$, which is what we wanted to show.

Whether or not we take satisfaction from the fact that f^{KM} is implementable, it is only an example. Clearly, what is needed is a set of general criteria for whether *any* given SCR is implementable. It is to this task to which I now turn.

IV. MONOTONICITY AND NO VETO POWER

The condition on SCRs that is central to their implementability is monotonicity.

Monotonicity.[12] The SCR $f: \mathscr{R} \to\!\!\!\!\to A$ satisfies monotonicity provided that $\forall a \in A$, $\forall \boldsymbol{R}, \boldsymbol{R}' \in \mathscr{R}$ if $a \in f(\boldsymbol{R})$ and $[\forall i \in \{1, \ldots, n\}\ \forall b \in A\ aR_i b \Rightarrow aR_i' b]$, then $a \in f(\boldsymbol{R}')$.

In words, monotonicity requires that if alternative a is f-optimal with respect to some profile of preferences and the profile is then altered so that, in each individual's ordering, a does not fall below any alternative that it was not below before, then a remains f-optimal with respect to the new profile.

To see that monotonicity is "reasonable", let us observe that it is satisfied by the prominent SCRs mentioned in Section II. First consider the Pareto correspondence f^{PO}. If a is (weakly) Pareto optimal with respect to $\boldsymbol{R}$ then, for all b, there exists j_* such that $aR_{j_*} b$. But if we replace $\boldsymbol{R}$ by $\boldsymbol{R}'$ such that, for all i, $aR_i b \Rightarrow aR_i' b$, we conclude that $aR_{j_*}' b$. Hence, b is (weakly) Pareto optimal with respect to $\boldsymbol{R}'$, establishing the monotonicity of f^{PO}.

Next, let us examine the Condorcet correspondence f^{CON}. If a is a majority winner for a strict profile (a profile consisting of strict orderings) $\boldsymbol{R}$, then, for any other alternative b, the number of individuals preferring a to b is no less than the number preferring b to a

$$\#\{i \mid aR_i b\} \geq \#\{i \mid bR_i a\}. \tag{3}$$

But if $\boldsymbol{R}'$ is a profile such that, for all i, $aR_i b \Rightarrow aR_i' b$, then the left-hand side of (3) cannot fall when we replace $\boldsymbol{R}$ by $\boldsymbol{R}'$. Furthermore, if the right-hand side rises, then we must have $aR_i b$ and $bR_i' a$ for some i, a contradiction of the relation between $\boldsymbol{R}$ and $\boldsymbol{R}'$, given the strictness of preferences. We conclude that (3) continues to hold when $\boldsymbol{R}'$ replaces $\boldsymbol{R}$, and so a is still a majority winner.

12. Monotonicity is called "strong positive association" by Muller and Satterthwaite (1977), who show that when f is *single-valued* and the domain consists of all strict preferences $\mathscr{R}^*$ then monotonicity is necessary and sufficient for implementation in dominant strategies. However, more generally, when f is either nonsingle-valued or the domain of preferences admits indifference or is more highly restricted than $\mathscr{R}^*$, this characterization result no longer obtains (see Dasgupta, Hammond, and Maskin (1979)).

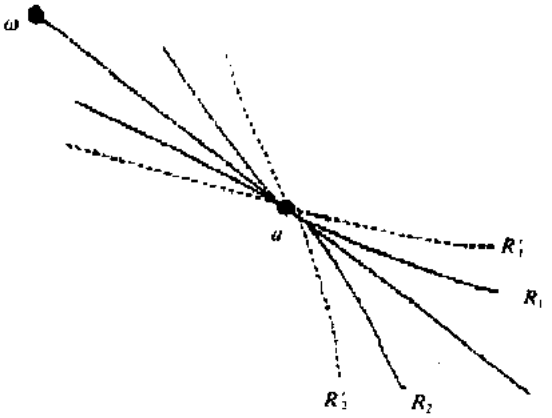

FIGURE 1
The Walrasian correspondence

As for the Walrasian correspondence, refer to the Edgeworth Box in Figure 1. In this two-consumer, two-good economy, allocation a is a competitive equilibrium allocation with respect to the endowments ω and the preference profile $R = (R_1, R_2)$. If we now alter R so that any allocation that was worse than a for consumer i remains worse than a, we obtain profile $R' = (R_1', R_2')$, with respect to which a remains a competitive equilibrium. Hence, f^W is monotonic.[13]

We should also point out that monotonicity is *automatically* satisfied by any SCR whose domain of preferences is among certain classes of preferences often studied in the literature. For example, this is true of classes of preferences satisfying the "single-crossing" property (*i.e.* the "Spence/Mirrlees" condition). A set of preferences $\mathcal{R}_i$ satisfies this property if no two indifference curves in the class intersect more than once. Notice that this means (refer to Figure 2) that if $R_i, R_i' \in \mathcal{R}_i$ and $a, b \in A$ are such that aR_ib and $aR_i'b$, then there exists another alternative $b' \in A$ such that aR_ib' but $b'P(R_i')a$. Hence, the hypothesis of the monotonicity condition cannot be satisfied, and so the condition holds vacuously.

A particularly interesting case in which single-crossing holds is that in which alternatives are nondegenerate lotteries over a set of possible outcomes. If individuals' preferences over lotteries satisfy the von Neumann–Morgenstern axioms, then indifference curves in probability space are straight, parallel lines, and so clearly indifference curves corresponding to distinct preferences can intersect only once. This insight figures prominently in the literature on "virtual implementation" (see Abreu and Sen (1991) and Abreu and Matsushita (1992)).

For an example of a well-known SCR that fails to satisfy monotonicity, consider the Borda Court (*i.e.* rank-order voting) SCR f^{BC}. For each individual, according to f^{BC}, points are assigned to each of the m alternatives available: m points are assigned to his favourite alternative, $m-1$ to his next favourite, and so on. The alternative (or alternatives) chosen by f^{BC} is the one for which the sum of points over individuals is highest. Suppose that $A = \{a, b, c, d\}$ (*i.e.* $m = 4$) and $n = 2$. Consider the profile $R = (R_1, R_2)$:

13. This argument relies on competitive allocations like a being *interior* allocations. For what can go wrong if a competitive allocation occurs on the boundary, see Hurwicz, Maskin, and Postlewaite (1995).

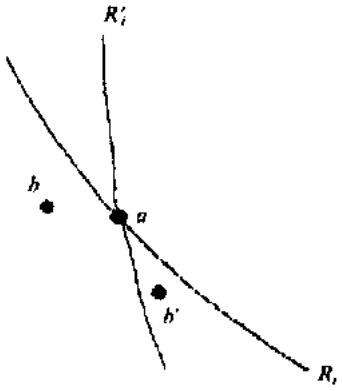

FIGURE 2
Single-crossing preferences

R_1	R_2
a	c
d	b
b	a
c	d

Note that in this profile, alternative a garners the most points (6), and so is chosen by f^{BC}. Next consider the profile $R' = (R_1', R_2')$:

R_1'	R_2'
a	b
b	c
d	a
c	d

Notice that in going from R_i to R_i', a does not fall *vis-à-vis* any other alternative. Thus, monotonicity would require that it still be chosen for profile R'. However, a no longer attracts the most points; alternative b does (7). Hence monotonicity is violated.

Whether one accepts monotonicity as natural or has qualms about its restrictiveness, it is an inescapable requirement for implementability in Nash equilibrium, as the following result shows.

Theorem 2. *If $f: \mathscr{R} \twoheadrightarrow A$ is an* SCR *that is implementable in Nash equilibrium, then it is monotonic.*

Proof. Suppose that f is implementable in Nash equilibrium by the game form $g: S_1 \times \cdots \times S_n \to A$. For some profile $R \in \mathscr{R}$ consider $a \in f(R)$. Then there exists a Nash equilibrium s of g with respect to R such that $g(s) = a$. Consider profile $R' \in \mathscr{R}$ such that

$$\text{for all } i \text{ and all } b \in A \qquad aR_i b \Rightarrow aR'_i b. \tag{4}$$

If there exist i and s'_i such that $g(s'_i, s_{-i})P(R'_i)g(s) = a$, then from (4), $g(s'_i, s_{-i})P(R_i)a$, a contradiction of the assumption that s is a Nash equilibrium with respect to R. Hence s is also a Nash equilibrium with respect to R'. From requirement (2) of the definition of implementability, therefore, we conclude that $a \in f(R')$. Thus f is monotonic. ||

I will show below (Theorem 3) that not only is monotonicity a necessary condition for implementability, as just demonstrated, but almost a sufficient condition as well. Nevertheless monotonicity by itself does not suffice to ensure implementability (see Example 2 below). One weak condition that we can add to monotonicity to do the trick is no veto power:

No Veto Power (NVP). An SCR $f: \mathscr{R} \to\!\!\!\!\to A$ satisfies NVP if, for all $R \in \mathscr{R}$ and all $a \in A$, whenever there exists $i \in \{1, \ldots, n\}$ such that, for all $j \neq i$ and all $b \in A$, $aR_j b$, then $a \in f(R)$.

NVP says that if an alternative is at the top of $n-1$ individuals' preference orderings, then the last individual cannot prevent the alternative from being f-optimal (*i.e.* he cannot "veto" it).

NVP is satisfied by virtually all "standard" SCRs (including the Pareto and Condorcet correspondences). It is also often automatically satisfied by any SCR when preferences are restricted. Consider, for example, a pure exchange economy with at least three consumers, in which an alternative corresponds to an allocation of goods. If there exists at least one good that is transferable and desirable, that gives rise to no externalities, and that is available in a positive quantity, then it is impossible to find an alternative that all but one consumer rank at the top of their preference orderings. This is because, for an individual to prefer a given allocation to all others, the allocation must assign him all of the good in question; if any other individuals got some of this good, he would be better off receiving their portions. Clearly, no other individual could also rank this allocation first, since it cannot be the case the two consumers each receive all of the good in question. Therefore the NVP property is satisfied vacuously.

V. SUFFICIENT CONDITIONS FOR IMPLEMENTATION

I now present the main result of the paper.

Theorem 3. *If $n \geq 3$ and $f: \mathscr{R} \to A$ is a n-person* SCR *satisfying monotonicity and* NVP, *then it is implementable in Nash equilibrium.*

Proof. The proof is by construction. I first show that we can construct a game form all of whose *pure-strategy* equilibria satisfy (1) and (2).[14] In the appendix I show that the construction can be extended to handle mixed strategies. For each player i, define the strategy space

$$S_i = \mathscr{R} \times A \times \{1, 2\},$$

14. This elegant proof is due essentially to Repullo (1987).

where $\mathcal{M}$ consists of the nonnegative integers. In other words, player i chooses as a strategy a triple consisting of a preference profile R^i in $\mathcal{R}$ (not necessarily the true one), an alternative a^i from A, and a number m^i in $\mathcal{M}$ (the numbers serve only to break ties).

For all i, all $R_i \in \mathcal{R}_i$ and all $a \in A$ define

$$L(a, R_i) = \{b | aR_i b\}.$$

$L(a, R_i)$ is the lower contour set of R_i at alternative a: the set of alternatives that are no better than a according to R_i.

I will construct an implementing game form g that implements f in three steps:

(i) If, for some $\boldsymbol{R}, a$, and m,

$$s_1 = \cdots = s_n = (\boldsymbol{R}, a, m) \text{ and } a \in f(\boldsymbol{R}), \qquad \text{take } g(s_1, \ldots, s_n) = a. \tag{5}$$

In words, if players are unanimous in their strategy, and their proposed alternative a is f-optimal given their proposed profile $\boldsymbol{R}$, the outcome is a.

(ii) If, for all $j \neq i$, $s_j = (\boldsymbol{R}, a, m)$, $s_i = (R^i, a^i, m^i) \neq (\boldsymbol{R}, a, m)$, and $a \in f(\boldsymbol{R})$ take

$$g(s_1, \ldots, s_n) = \begin{cases} a^i, & \text{if } a^i \in L(a, R_i) \\ a, & \text{if } a^i \notin L(a, R_i). \end{cases}$$

That is, suppose that all players but one play the same strategy and, given their proposed profile, their proposed alternative a is f-optimal. Then, the odd-man-out, gets his proposed alternative, provided that it is in the lower contour set at a of the ordering that the *other* players propose for him; otherwise, the outcome is a.

(iii) If neither (i) nor (ii) applies, then

$$g(s_1, \ldots, s_n) = a^{i^*}, \tag{6}$$

where $i^* = \max \{i | m^i = \max_j m^j\}$. In other words, when neither (i) nor (ii) applies, the outcome is the alternative proposed by the player with the highest index among those whose proposed number is maximal.

It remains to show that this game form implements f. I first claim that, for all $\boldsymbol{R} \in \mathcal{R}$ and all $a \in A$, if $a \in f(\boldsymbol{R})$, then, for any $m \in \mathcal{M}$, the strategy profile $(s_1, \ldots, s_n)$ satisfying (5) constitutes a Nash equilibrium with respect to $\boldsymbol{R}$. To see this, note from (i) that the outcome from this strategy profile is a. Moreover, from (ii), any player i who deviates unilaterally from $(s_1, \ldots, s_n)$ gets an alternative in $L(a, R_i)$, which, by definition of the lower contour set, is no better for him than a. Thus I have established requirement (1)—that there is a Nash equilibrium of g corresponding to each f-optimal alternative—in the definition of implementability.

To establish (2)—that every Nash equilibrium of g is f-optimal—consider first a Nash equilibrium $(s_1, \ldots, s_n)$ in which (5) holds and $a \in f(\boldsymbol{R})$, but where the true preference profile is $\boldsymbol{R}'$. From (i), the equilibrium outcome is a. Moreover, because $(s_1, \ldots, s_n)$ is an equilibrium with respect to $\boldsymbol{R}'$, (ii) implies that

$$\text{for all } i \text{ and all } b \in L(a, R_i), \qquad aR_i' b. \tag{7}$$

(To understand why (7) holds, note that if instead we had $bP(R_i')a$ for some i and $b \in L(a, R_i)$, it would pay player i to deviate from s_i and induce b, which (ii) implies he could do. But this would contradict the assumption that $(s_1, \ldots, s_n)$ is an equilibrium.) But (7) can be rewritten as

$$\text{for all } i \text{ and all } b \in A, \; aR_i b \Rightarrow aR_i' b. \tag{8}$$

Hence because f satisfies monotonicity, (8) and the fact that $a \in f(R)$ imply that $a \in f(R')$, *i.e.* the equilibrium outcome is f-optimal.

Next let us consider a Nash equilibrium $(s_1, \ldots, s_n)$ for R' in which, for all $j \neq i$,

$$s_j = (R, a, m),$$

where $a \in f(R)$, but $s_i \neq (R, a, m)$, *i.e.* the strategy profile is such that (ii) applies. Let the outcome from this profile be a'. From (iii), each player $j \neq i$ could deviate from s_j and induce any alternative $a \in A$ he wishes by choosing m^j high enough (*i.e.* higher than $\max_{k \neq j} m^k$). Hence, the fact that $(s_1, \ldots, s_n)$ is a Nash equilibrium for R' implies that, for all $j \neq i$,

$$a' R'_j b \quad \text{for all } b \in A. \tag{9}$$

We conclude that NVP together with (9) ensures that $a' \in f(R')$, *i.e.* the equilibrium outcome is again f-optimal.

The same argument as in the preceding paragraph applies if $(s_1, \ldots, s_n)$ is a Nash equilibrium for which (iii) applies. ||

Remark 1. The game form constructed in the proof of Theorem 3 may be considered rather complicated. However, much of the complexity derives from its generality—the fact that it is supposed to work for a vast array of possible SCRs. For a *specific* SCR, by contrast, it is often possible to find an implementation that is quite simple (*e.g.* the mechanism in Example 1).

Remark 2. Even if the set of alternatives A is finite, the game form constructed in the proof of Theorem 3 has an unbounded strategy space, since $\mathbb{N}$ is unbounded. Jackson (1992) points out, however, that, for some solution concepts, the set of SCRs implementable by unbounded game forms is strictly larger than the limit of those implementable by bounded game forms as the bound goes to infinity. It remains an open question whether this is so for Nash equilibrium.

We have argued that no veto power is a weak condition. It is nevertheless restrictive, and so it is of some interest understanding its role in Theorem 3.[15] As we will see below (Theorem 4) NVP is not necessary for implementability. However, as the following example establishes, monotonicity by itself does not suffice.

Example 2. Suppose that $n = 3$ and $A = \{a, b, c\}$. For all i, let $\mathscr{R}_i = \mathscr{R}^*_A$ (*i.e.* the domain consists of all strict orderings). Define f^* such that, for all $R \in \mathscr{R}$,

for each $x \in \{a, b\}$, $x \in f^*(R)$ if and only if x is Pareto-optimal and top-ranked for individual 1

$c \in f^*(R)$ if and only if c is Pareto optimal and not bottom-ranked for individual 1.

It is easy to verify that f^* is monotonic. However, it does not satisfy NVP because if individual 1 bottom-ranks alternative c, it fails to be f^*-optimal even if individuals 2 and 3 top-rank c.

15. For conditions that are necessary and sufficient for implementability, see Moore and Repullo (1990).

Consider the following three profiles R^*, R^{**}, R^{***}:

$$R^* = ([b, c, a], [c, a, b], [c, a, b])$$
$$R^{**} = ([a, b, c], [c, b, a], [c, a, b])$$
$$R^{***} = ([b, a, c], [a, b, c], [a, b, c]),$$

where "$[x, y, z]$" denotes the ordering in which x is preferred to y, and y is preferred to z. Then

$$f^*(R^*) = \{b, c\}, f^*(R^{**}) = \{a\}, f^*(R^{***}) = \{b\}.$$

If f^* were implementable, there would exist a game form g and a Nash equilibrium $s^* = (s_1^*, s_2^*, s_3^*)$ with respect to R^* such that $g(s^*) = c$. Because $bP(R_1^*)c$ there does not exist $s_1'' \in S_1$ such that $g(s_1'', s_2^*, s_3^*) = b$.

If there existed $s_1' \in S_1$ such that $g(s_1', s_2^*, s_3^*) = a$, then (s_1', s_2^*, s_3^*) would be a Nash equilibrium for R^{***}, a contradiction since $a \notin f(R^{***})$. Hence, s_1' cannot exist. We conclude that (s_1^*, s_2^*, s_3^*) is a Nash equilibrium for R^{**}, which contradicts the fact that $c \notin f(R^{**})$. Hence, f^* is not implementable.

VI. INDIVIDUAL RATIONALITY

We will say that an SCR $f: \mathscr{R} \to\!\!\!\!\to A$ is *individually rational* (IR) with respect to some alternative $a^0 \in A$ if for all $R \in \mathscr{R}$, all $a \in f(R)$, and all i, $aR_i a^0$. That is, if a is f-optimal, it must be weakly preferred by all individuals to a^0. In general, an SCR satisfying IR does not satisfy NVP because if, for some profile preference, everyone but individual i top-ranks alternative a, then NVP would require that a be f-optimal with respect to that profile. But f would then violate IR if i strictly preferred a^0 to a.

Nevertheless, many SCRs satisfying IR are implementable. One example is the "Individual Rationality" correspondence: for all $R \in \mathscr{R}$

$$f^{\mathrm{IR}}(R) = \{a \in A \mid aR_i a^0 \text{ for all } i\}.$$

This SCR is implemented by the game form g^{IR} such that $S_i = A$ for all i and

$$g^{\mathrm{IR}}(s_1, \ldots, s_n) = \begin{cases} a, & \text{if } s_1 = \cdots = s_n = a \text{ for some } a \in A \\ a^0, & \text{otherwise.} \end{cases}$$

The example of f^{IR} suggests that if we relax NVP so that it applies only to individually rational alternatives, we might obtain a general result.

Weak no veto power (WNVP). An SCR $f: \mathscr{R} \to\!\!\!\!\to A$ satisfies WNVP with respect to a^0 if, for all $R \in \mathscr{R}$ and all $a \in A$, whenever there exists i such that for all $j \neq i$ $aR_j b$ for all b and $aR_i a^0$, then $a \in f(R)$.

Theorem 4. *If $n \geq 3$ and $f: \mathscr{R} \to\!\!\!\!\to A$ is an* SCR *satisfying monotonicity,* WNVP, *and* IR *with respect to a^0, then it is implementable in Nash equilibrium.*

Proof. The proof of Theorem 4 uses exactly the same construction as that of Theorem 3. The only thing to show is that Nash equilibria satisfying configurations (ii) or (iii) in the proof of Theorem 3 are individually rational. This enables us to apply WNVP and complete the proof.

Thus, consider a configuration (ii) equilibrium $(s_1, \ldots, s_n)$ with respect to profile R'. That is, there exist i, a, m, and R such that $a \in f(R)$ and, for all $j \neq i$, $s_j = (R, a, m)$. We must show that $g(s_1, \ldots, s_n) R_k a^0$ for all k. This is immediate for all $j \neq i$, since each of those players can deviate from s_j and induce his top-ranked alternative. Hence, the fact that $(s_1, \ldots, s_n)$ constitutes an equilibrium means that $g(s_1, \ldots, s_n)$ itself must be top-ranked and so $g(s_1, \ldots, s_n) R'_j a^0$. As for player i, note that because f satisfies IR with respect to a^0, $a R_i a^0$, *i.e.* $a^0 \in L(R_i, a)$. Hence, by construction of g, player i can induce a^0 when the other players all use strategy (R, a, m). Thus because $(s_1, \ldots, s_n)$ is an equilibrium implies that $g(s_1, \ldots, s_n) R'_i a^0$. The argument is virtually identical for configuration (iii). ‖

APPENDIX

Proof of Theorem 3 for mixed strategies

The argument that all Nash equilibria of the mechanism in Theorem 3 are f-optimal does not carry over to mixed strategies.

To see the problem consider a mixed-strategy equilibrium $(\mu_1, \ldots, \mu_n)$ (for profile R') for which one possible realization is $s = (s_1, \ldots, s_n)$ where, for some j, $R \in \mathscr{R}$, and $a \in f(R)$,

$$s_i = (R, a, 0) \quad \text{for all } i \neq j,$$

but $s_j \neq (R, a, 0)$. In the proof of Theorem 3, I showed that the outcome corresponding to s must be f-optimal since, by deviating from s_i, each player $i \neq j$ could induce his favourite alternative a^i (NVP then would imply f-optimality of the outcome). But if there are other possible realizations of μ_{-i}, then player i might suffer by trying to induce a^i. Suppose, for example, that s'_{-i} is a realization in which, for some $R' \in \mathscr{R}$ and $a' \in f(R')$

$$s'_k = (R', a', 0) \quad \text{for all } k \neq i.$$

Assume, furthermore, that

$$a^i P(R'_i) a' \tag{*}$$

Then, although individual i can induce a^i against s_{-i}, formula (*) and the construction of Theorem 3 imply that he cannot induce a^i against s'_{-i}. Indeed, if he tries to do so, the outcome will be a', which may be a very bad alternative for him.

I now show, however, that the game form from Theorem 3 can be modified to circumvent this difficulty. For each player i, define the strategy space

$$S_i = \mathscr{R} \times A \times \{\alpha \mid \alpha : \mathscr{R}^n \times A^n \to A\} \times \{1, \ldots, n\}.$$

In other words, player i chooses as a strategy a quadruple consisting of a preference profile $R' \in \mathscr{R}$, an alternative $a' \in A$, a function $\alpha'(\cdot)$ mapping each possible vector of announced profiles and alternatives $(R^1, \ldots, R^n, a^1, \ldots, a^n)$ to an alternative $\alpha'(R^1, \ldots, R^n, a^1, \ldots, a^n) \in A$, and an integer $m' \in \{1, \ldots, n\}$.

As in the proof of Theorem 3, I will construct the implementing game form g in three steps:

(i) If $s_1 = \cdots = s_n = (R, a, \alpha(\cdot), m)$ and $\alpha(R, \ldots, R, a, \ldots, a) = a \in f(R)$, take $g(s_1, \ldots, s_n) = a$.

In other words, if players are unanimous in their strategy, and their proposed alternative a is prescribed by their proposed function $\alpha(\cdot)$ and f-optimal given their proposed profile R, the outcome is a.

(ii) If, for all $j \neq i$, $s_j = (R, a, \alpha(\cdot), m)$ with $\alpha(R, \ldots, R, a, \ldots, a) = a \in f(R)$ but $s_i = (R^i, a^i, \alpha^i(\cdot), m^i) \neq (R, a, \alpha(\cdot), m)$, take

$$g(s_1, \ldots, s_n) = \begin{cases} \alpha^i(R, \ldots, R^i, \ldots, R, a, \ldots, a^i, \ldots, a), & \text{if } \alpha^i(R, \ldots, R^i, \ldots, R, a, \ldots, a^i, \ldots, a) \in L(a, R_i), \\ a, & \text{otherwise.} \end{cases}$$

That is, suppose that all players but player i play the same strategy and their proposed alternative a is prescribed by their proposed function $\alpha(\cdot)$ and f-optimal, given their proposed profile R. Then, player i gets the alternative prescribed by his proposed function $\alpha^i(\cdot)$ (given the vector of proposed profiles and alternatives

$(R, \ldots, R^i, \ldots, R, a, \ldots, a^i, \ldots, a))$ provided that it is in the lower contour set at a of the ordering that the other players propose for him; otherwise, the outcome is a.

(iii) If neither (i) nor (ii) applies, then

$$g(s_1, \ldots, s_n) = \alpha^{i^*}(R^1, \ldots, R^n, a^1, \ldots, a^n), \tag{A1}$$

where $i^* = \max\{i | m^i = \max_j m^j\}$. That is, the outcome is the alternative prescribed by the proposed function of the player whose index is highest among those proposing the maximal number.

I must show that this game form implements f. Note first that, for all $R \in \mathscr{R}$ and all $a \in A$, if $a \in f(R)$, then, for any $m \in \mathscr{M}$, the strategy profile $(s_1, \ldots, s_n)$ where

$$s_1 = \cdots = s_n = (R, a, \alpha(\cdot), m) \quad \text{and} \quad \alpha(R, \ldots, R, a, \ldots, a) = a, \tag{A2}$$

constitutes a Nash equilibrium with respect to R. To see this, note from (i) that the outcome from this strategy profile is a. Moreover, from (ii), any player i who deviates unilaterally from $(s_1, \ldots, s_n)$ induces an alternative in $L(a, R_i)$, which by definition of $L(a, R_i)$ is no better for player i (with preference ordering R_i). Thus the profile (A2) indeed constitutes a Nash equilibrium with respect to R, and so I have established that, for every f-optimal alternative, there is a Nash equilibrium of g giving rise to that alternative.

It remains to show that if $(\mu_1, \ldots, \mu_n)$ is a Nash equilibrium for g with respect to profile R', then the outcome corresponding to each realization $(s_1^*, \ldots, s_n^*)$ in the support of $(\mu_1, \ldots, \mu_n)$ is f-optimal. Suppose first that $(s_1^*, \ldots, s_n^*)$ is a realization for which (A2) holds and $a \in f(R)$, but that the profile R differs from the true profile R'. From (i), the equilibrium outcome is a. For any player i, consider $b \in A$ such that $aR_i b$. Now imagine that player i plays $s_i = (R', a', \alpha', m')$ such that

$$(R', a', m') = (R, a, m), \tag{A3}$$

and

$$\alpha^i(\hat{R}^1, \ldots, \hat{R}^n, \hat{a}^1, \ldots, \hat{a}^n) = \begin{cases} b, & \text{if } (\hat{R}^1, \ldots, \hat{R}^n, \hat{a}^1, \ldots, \hat{a}^n) = (R, \ldots, R, a, \ldots, a), \\ \alpha(\hat{R}^1, \ldots, \hat{R}^n, \hat{a}^1, \ldots, \hat{a}^n), & \text{otherwise.} \end{cases} \tag{A4}$$

That is, s_i is the same as $s_i^* = (R, a, \alpha(\cdot), m)$ except for the function $\alpha^i(\cdot)$, which, in turn, is the same as $\alpha(\cdot)$ except at the point $(R, \ldots, R, a, \ldots, a)$. Now because $b \in L(a, R_i)$, (ii), (A3), and (A4) together imply that

$$g(s_i, s_{-i}^*) = b. \tag{A5}$$

Moreover, because $\alpha^i(\cdot)$ is the same as $\alpha(\cdot)$ except at $(R, \ldots, R, a, \ldots, a)$,

$$g(s_i, \mathfrak{s}_{-i}) = g(s_i^*, \mathfrak{s}_{-i}) \tag{A6}$$

for any realization $\mathfrak{s}_{-i} = (\hat{R}^{-i}, \hat{a}^{-i}, \hat{\alpha}^{-i}(\cdot), \hat{m}^{-i})$ of μ_{-i} such that $(\hat{R}^{-i}, \hat{a}^{-i}) \neq (R, \ldots, R, a, \ldots, a)$. Hence, if $bP(R_i')a$, (A5) and (A6) imply that player i is better off using s_i rather than s_i^* against μ_{-i}. We conclude that

$$\text{for all } i \text{ and all } b \qquad aR_i b \Rightarrow aR_i' b. \tag{A7}$$

Hence, because f is monotonic (A7) and the fact that $a \in f(R)$ imply that $a \in f(R')$. That is, the outcome $g(s_1^*, \ldots, s_n^*)$ is f-optimal, as required.

Next let us consider a realization $(s_1^*, \ldots, s_n^*)$ in the support of $(\mu_1, \ldots, \mu_n)$ in which, for all $j \neq i$,

$$s_j^* = (R, a, \alpha(\cdot), m),$$

where $\alpha(R, \ldots, R, a, \ldots, a) = a \in f(R)$ but $s_i^* = (R', a', \alpha^i(\cdot), m^i) \neq (R, a, \alpha(\cdot), m)$. That is, the strategy profile is such that (ii) applies. Let the outcome from this profile be a'. For any $j \neq i$, choose $b^j \in A$ such that

$$b^j R_j' b \quad \text{for all } b \in A. \tag{A8}$$

Then, consider $s_j = (R^j, a^j, \alpha^j(\cdot), m^j)$ such that

$$(R^j, a^j) = (R, a), \tag{A9}$$

$$m^j > \max\{m^i, m\}, \tag{A10}$$

and

$$\alpha^j(\hat{R}^1, \ldots, \hat{R}^n, \hat{a}^1, \ldots, \hat{a}^n) = \begin{cases} b^j, & \text{if } (\hat{R}^1, \ldots, \hat{R}^n, \hat{a}^1, \ldots, \hat{a}^n) = (R, \ldots, R^i, \ldots, R, a, \ldots, a^i, \ldots, a), \\ \alpha(\hat{R}^1, \ldots, \hat{R}^n, \hat{a}^1, \ldots, \hat{a}^n), & \text{otherwise.} \end{cases} \tag{A11}$$

That is, s_j is the same as $s_j^* = (R, a, \alpha(\cdot), m)$ except for the integer m_j (which we take to be bigger than the other numbers in s_{-j}^*) and the function $\alpha^j(\cdot)$, which, in turn, is the same as $\alpha(\cdot)$ except at the point $(R, \ldots, R^i, \ldots, R, a, \ldots, a^i, \ldots, a)$. From (A9)–(A11),

$$g(s_j, s_{-j}^*) = b^i. \tag{A12}$$

Moreover, for each $\hat{s}_{-j} \neq s_{-j}^*$ (A9)-(A11) ensure that either

$$g(s_i, \hat{s}_{-i}) = b^j, \tag{A13}$$

or

$$g(s_j, \hat{s}_{-j}) = g(s_j^*, \hat{s}_{-j}). \tag{A14}$$

Hence, from (A8) and (A12)–(A14), we conclude that player j does strictly better with s_j than with s_j^* against μ_{-j}, a contradiction, unless player j does not strictly prefer b^i to a^i, *i.e.* unless

$$a^i R_j^i b \quad \text{for all } b \in A. \tag{A15}$$

Thus (A15) must hold for all $j \neq i$, and so, from NVP, $a^i \in f(R^i)$, as required.

The same argument as in the preceding paragraph applies if $(s_1, \ldots, s_n)$ is a realization in the support of $(\mu_1, \ldots, \mu_n)$ to which (iii) applies. ||

Remark. The reason for having players report functions $\alpha^i(\cdot)$ rather than merely fixed alternatives is to be able to accommodate mixed strategies. Which alternative is best for a player to propose will, in general, depend on the profiles and alternatives that the other players propose. But if the others are playing mixed strategies, then a player may not be able to forecast (except probabilitically) what these proposals will be. Allowing him to propose a function enables him, in effect, to propose an alternative on a contingent basis.

Acknowledgements. This work was supported by a grant from the National Science Foundation and a research fellowship from Jesus College, Cambridge. The paper was first presented at the Summer Workshop of the Econometric Society in Paris, June, 1977.

For useful conversations about the original 1977 version I thank Frank Hahn, Peter Hammond, Jerry Hausman, Leo Hurwicz, Roy Radner, and William Thomson. I am especially indebted to Karl Vind, whose suggestions led to a substantial strengthening and simplification of my results. I thank a referee for comments on the 1998 version.

I am most grateful to the Editor, Patrick Bolton, for giving me the opportunity to publish this elderly paper. The literature on implementation is now very large. But because there are a number of excellent recent surveys (see Moore (1993), Palfrey (1993) and (1998), Chapter 10 of Osborne and Rubinstein (1994), and Corchon (1996); see also my old survey Maskin (1985)), I have not made a systematic attempt to bring the references up to date. Indeed, I have made as few changes as possible to the 1977 text.

REFERENCES

ABREU, D. and SEN, A. (1991), "Virtual Implementation in Nash Equilibrium", *Econometrica*, **59**, 997–1021.

ABREU, D. and MATSUSHIMA, H. (1992), "Virtual Implementation in Iteratively Undominated Strategies: Complete Information", *Econometrica*, **60**, 993–1008.

ARROW, K. (1951) *Social Choice and Individual Values* (New York: John Wiley).

CORCHON, L. (1996) *The Theory of Implementation of Socially Optimal Decisions in Economics* (New York: St. Martin's Press).

DASGUPTA, P., HAMMOND, P. and MASKIN, E. (1979), "The Implementation of Social Choice Rules", *Review of Economic Studies*, **46**, 185–216.

DUTTA, B. and SEN, A. (1991), "A Necessary and Sufficient Condition for Two-Person Nash Implementation", *Review of Economic Studies*, **58**, 121–128.

GIBBARD, A. (1973), "Manipulation of Voting Schemes: A General Result", *Econometrica*, **41**, 587–602.

GROVES, T. and LEDYARD, J. (1977), "Optimal Allocation of Public Goods: A Solution to the Free Rider Problem", *Econometrica*, **41**, 617–631.

HURWICZ, L. (1979*a*), "Outcome Functions Yielding Walrasian and Lindahl Allocations at Nash Equilibrium", *Review of Economic Studies*, **46**, 217–225.

HURWICZ, L. (1979*b*), "On Allocations Attainable Through Nash Equilibria", *Journal of Economic Theory*, **21**, 140–165.

HURWICZ, L., MASKIN, E. and POSTLEWAITE, A. (1995), "Feasible Nash Implementation of Social Choice Correspondences when the Designer Does Not Know Endowments or Production Sets", in J. Ledyard (ed.), *The Economics of Informational Decentralization: Complexity, Efficiency and Stability* (Dordrecht: Kluwer Academic Publishers).

HURWICZ, L. and SCHMEIDLER, D. (1978), "Outcome Functions which Guarantee the Existence and Pareto Optimality of Nash Equilibria", *Econometrica*, **46**, 144–174.

JACKSON, M. (1992), "Implementation in Undominated Strategies: A Look at Bounded Mechanisms", *Review of Economic Studies*, **59**, 757–775.

MASKIN, E. (1985), "The Theory of Nash Equilibrium: A Survey", in L. Hurwicz, D. Schmeidler, and H. Sonnenschein, *Social Goals and Social Organization* (Cambridge: Cambridge University Press), 173–204.

MOORE, J. (1993), "Implementation, Contracts, and Renegotiation in Environments with Complete Information", in J.-J. Laffont (ed.), *Advances in Economic Theory* (Cambridge: Cambridge University Press), 182–282.

MOORE, J. and REPULLO, R. (1990), "Nash Implementation: A Full Characterization", *Econometrica*, **58**, 1083–1099.

MULLER, E. and SATTERTHWAITE, M. (1977), "The Equivalence of Strong Positive Association and Strategy Proofness", *Journal of Economic Theory*, **14**, 412–418.

OSBORNE, M. and RUBINSTEIN, A. (1994) *A Course in Game Theory* (Cambridge: MIT Press).

PALFREY, T. (1993), "Implementation in Bayesian Equilibrium: The Multiple Equilibrium Problem in Mechanism Design", in J.-J. Laffont (ed.), *Advances in Economic Theory* (Cambridge: Cambridge University Press).

PALFREY, T. (1998), "Implementation Theory", forthcoming in R. Aumann and S. Hart (eds.), *Handbook of Game Theory, Vol. 3* (Amsterdam: North-Holland).

REPULLO, R. (1987), "A Simple Proof of Maskin's Theorem on Nash Implementation", *Social Choice and Welfare*, **4**, 39–41.

SATTERTHWAITE, M. (1975), "Strategy-Proofness and Arrow's Conditions: Existence and Correspondence Theorems for Voting Procedures and Social Welfare Functions", *Journal of Economic Theory*, **10**, 187–217.

SCHMEIDLER, D. (1980), "Walrasian Analysis via Strategic Outcome Functions", *Econometrica*, **48**, 1585–1595.

□ □ □

THE FOLK THEOREM IN REPEATED GAMES WITH DISCOUNTING OR WITH INCOMPLETE INFORMATION[1]

BY DREW FUDENBERG AND ERIC MASKIN[2]

When either there are only two players or a "full dimensionality" condition holds, any individually rational payoff vector of a one-shot game of complete information can arise in a perfect equilibrium of the infinitely-repeated game if players are sufficiently patient. In contrast to earlier work, mixed strategies are allowed in determining the individually rational payoffs (even when only realized actions are observable). Any individually rational payoffs of a one-shot game can be approximated by sequential equilibrium payoffs of a long but finite game of incomplete information, where players' payoffs are almost certainly as in the one-shot game.

1. INTRODUCTION

THAT STRATEGIC RIVALRY in a long-term relationship may differ from that of a one-shot game is by now quite a familiar idea. Repeated play allows players to respond to each other's actions, and so each player must consider the reactions of his opponents in making his decision. The fear of retaliation may thus lead to outcomes that otherwise would not occur. The most dramatic expression of this phenomenon is the celebrated "Folk Theorem" for repeated games. An outcome that Pareto dominates the minimax point is called individually rational. The Folk Theorem asserts that any individually rational outcome can arise as a Nash equilibrium in infinitely repeated games with sufficiently little discounting. As Aumann and Shapley [3] and Rubinstein [20] have shown, the same result is true when we replace the word "Nash" by "(subgame) perfect" and assume no discounting at all.

Because the Aumann–Shapley/Rubinstein result supposes literally no discounting, one may wonder whether the *exact* counterpart of the Folk Theorem holds for perfect equilibrium, i.e., whether as the discount factor tends to one, the set of perfect equilibrium outcomes converges to the individually rational set. After all, agents in most games of economic interest are not completely patient; the no discounting case is of interest as an approximation.

It turns out that this counterpart is false. There can be a discontinuity (formally, a failure of lower hemicontinuity) where the discount factor, δ, equals one, as we show in Example 3. Nonetheless the games in which discontinuities occur are quite degenerate, and, in the end, we *can* give a qualified "yes" (Theorem 2) to the question of whether the Folk Theorem holds with discounting. In particular, it always holds in two-player games (Theorem 1). This last result contrasts with the recent work of Radner–Myerson–Maskin [18] showing that, even in two-player games, the equilibrium set may not be continuous at $\delta = 1$ in

[1] Under stronger hypotheses, we obtain a sharper characterization of perfect and Nash equilibrium payoffs of discounted repeated games in our note (Fudenberg and Maskin [8]).

[2] We wish to thank D. Abreu, R. Aumann, D. Kreps, M. Whinston, and three referees for helpful comments. The project was inspired by conversations with P. Milgrom. We are grateful to NSF Grants SES 8409877 and SES 8320334 and the Sloan Foundation for financial support.

the discount factor if players' moves are not directly observable and outcomes depend stochastically on moves.

Until recently, the study of perfect equilibrium in repeated games concentrated mainly on infinite repetitions without discounting ("supergames"). One early exception was Friedman [5 and 6], who showed that any outcome that Pareto dominates a Nash equilibrium of the constituent game (the game being repeated) can be supported in a perfect equilibrium of the repeated game.[3] The repeated game strategies he specified are particularly simple: after any deviation from the actions that sustain the desired outcome, players revert to the one-shot Nash equilibrium for the remainder of the game. More recently, Abreu [1] established that a highly restricted set of strategies suffices to sustain any perfect equilibrium outcome. Specifically, whenever any player deviates from the desired equilibrium path, that player can be "punished" by players' switching to the worst possible equilibrium for the deviator regardless of the history of the game to that point.

We exploit this idea of history-free punishments, by contrast with the methods of Aumann-Shapley/Rubinstein, in the proofs of our Theorems 1 and 2, which are constructive.[4] In the proof of the two-person "discounting folk theorem" (Theorem 1), *both* players switch for a specified number of periods to strategies that minimize their opponent's maximum payoff (i.e., minimax strategies) after any deviation. Theorem 2 treats the n-person case, where "mutual minimaxing" may be impossible. In this case we impose a "full dimensionality" assumption that enables players to be rewarded for having carried out punishments. Theorem 5 also makes use of rewards to punishers to show that mixed strategies can be used as punishments even if only realized actions, and not the mixed strategies themselves, are observable. This provides a substantially stronger result, because the individually rational payoff levels are often lower with mixed strategies than with pure ones.

Although the theory of infinitely repeated games offers an explanation of cooperation in ongoing relationships, economic agents often have finite lives. If the game has a long but finite length, the set of equilibria may be much smaller than the folk theorem would suggest. The classic example here is the repeated prisoner's dilemma: with a fixed finite horizon the only equilibrium involves both players' confessing every period, in contrast with the cooperative equilibrium that is sustainable with an infinite horizon.[5] Still anecdotal and experimental evidence both suggest that cooperation is a likely outcome with a large but finite number of repetitions.

Recently Kreps-Wilson [14], Milgrom-Roberts [17], and Kreps-Milgrom-Roberts-Wilson [13] have proposed a reason why a finite number of repetitions might allow cooperation. Their explanation supposes that players are uncertain about the payoffs or possible actions of their opponents. Such "incomplete

[3] Actually Friedman was concerned explicitly only with Nash equilibria of the repeated game. The strategies that he proposed, however, constitute perfect equilibria (See Theorem C of Section 2).

[4] Lockwood [16] characterizes the (smaller set of) equilibrium payoffs that are possible when one restricts attention to punishments of the Aumann-Shapley/Rubinstein variety.

[5] See, however, Benoit and Krishna [4] and Friedman [7] who show that when a game with multiple equilibria is repeated even only finitely many times, "Folk-Theorem-like" results may emerge.

information" in the prisoner's dilemma precludes applying the backwards-induction argument that establishes that the players must confess each period. Players can credibly threaten to take suboptimal actions if there is some (small) probability that the action is indeed optimal, because they have an interest in maintaining their reputation for possible "irrationality."

The examples of reputation games analyzed to date exhibit the apparent advantage, compared with infinite-horizon models, of having substantially smaller sets of equilibria. However, the equilibrium set depends on the precise form of irrationality specified. Our "incomplete information" Folk Theorem shows that by varying the kind of irrationality specified, but still keeping the probability of irrationality arbitrarily small, one can trace out the entire set of infinite-horizon equilibria. Thus, in a formal sense, the two approaches, infinite and finite horizon, yield the same results. However, those who are willing to choose among different forms of irrationality may still find the incomplete information approach useful. One may argue for or against certain equilibria on the basis of the type of irrationality needed to support them.

We provide two different theorems for repeated games of incomplete information. Our first result (Theorem 3) parallels Friedman's work on repeated games with discounting: after a deviation the "crazy" player switches to a Nash-equilibrium strategy of the constituent game. This simple form of irrationality suffices to support any outcome that Pareto-dominates a (one-shot) Nash equilibrium. Our second, and main, result (Theorem 4) uses a more complex form of irrationality. However, the basic approach is the same as in our Folk Theorem with discounting: after a deviation each player switches to his minimax strategy for a specified number of periods.

It is not surprising that similar kinds of arguments should apply to both infinite horizon games with discounting and finite horizon games. Each type of game entails the difficulty, not present in infinite horizon games without discounting, that deviators from the equilibrium path cannot be "punished" arbitrarily severely. This limitation is a problem because of the requirement of perfection. Deviators must be punished, but it must also be in the interest of the punishers to punish. That is, they must themselves be threatened with punishment if they fail to punish a deviator. Such considerations give rise to an infinite sequence of potential punishments that, at each level, enforce the punishments of the previous level. Depending on how these punishments are arranged, they may have to become increasingly severe the farther out in the sequence they lie. This creates no problem in supergames but may be impossible for the two types of games that we consider. It seems natural, therefore, to study these two types together.

Section 2 presents the classical Folk Theorem and the Aumann-Shapley/Rubinstein and Friedman variants. Section 3 discusses continuity of the equilibrium correspondence as a function of the discount factor and develops Folk Theorems for infinitely repeated games with discounting. Section 4 provides a simple proof that any payoffs that Pareto dominate a (one-shot) Nash equilibrium can be sustained in an equilibrium of a finitely repeated game with incomplete information. This result is the analog of the Friedman [5] result. Section 5 uses a more complex approach to prove a Folk Theorem for these finitely repeated games.

Sections 2–5 follow previous work on repeated games in assuming that, if mixed strategies are used as punishments, they are themselves observable. In Section 6 we drop this assumption but show that our results continue to hold under the more natural hypothesis that players can observe only each other's past *actions*.

2. THE CLASSICAL FOLK THEOREM

Consider a finite n-person game in normal form

$$g: A_1 \times \cdots \times A_n \to R^n.$$

For now, we shall not distinguish between pure and mixed strategies, and so we might as well suppose that the A_i's consist of mixed strategies. Thus, we are assuming either that a player can observe the others' past mixed strategies, as in previous work on repeated games, or restricting players to pure strategies. Mixed strategies can be made observable if the outcomes of players' randomizing devices are jointly observable *ex-post*. (More importantly, we show in Section 6 that the assumption is not necessary.) Moreover, for convenience, we assume that the players can make their actions contingent on the outcome of a *public* randomizing device. That is, they can play correlated strategies.[6] Even if a correlated strategy over vectors of actions cannot literally be adopted, it can still be approximated if the action vectors are played successively over time and the frequency of any given vector corresponds to its probability in the correlated strategy. To see how to modify the statements of the theorems if correlated strategies cannot be used, see the Remark following Theorem A.

For each j, choose $M^j = (M^j_1, \ldots, M^j_n)$ so that

$$(M^j_1, \ldots, M^j_{j-1}, M^j_{j+1}, \ldots, M^j_n) \in \arg\min_{a_{-j}} \max_{a_j} g_j(a_j, a_{-j}),$$

and

$$v^*_j \equiv \max_{a_j} g_j(a_j, M^j_{-j}) = g_j(M^j).^7$$

The strategies $(M^j_1, \ldots, M^j_{j-1}, M^j_{j+1}, \ldots, M^j_n)$ are minimax strategies (which may not be unique) against player j, and v^*_j is the smallest payoff that the other players can keep player j below.[8] We will call v^*_j player j's reservation value and refer to $(v^*_1, \ldots, v^*_n)$ as the minimax point. Clearly, in any equilibrium of g—whether or not g is repeated—player j's expected average payoff must be at least v^*_j.

[6] See Aumann [2]. More generally, a correlated strategy might entail having each player make his action contingent on a (private) signal correlated with some randomizing device. We shall, however, ignore this possibility.

[7] The notation "a_{-j}" denotes "$(a_1, \ldots, a_{j-1}, a_{j+1}, \ldots, a_n)$", and "$g_j(a_j, M^j_{-j})$" denotes "$g_j(M^j_1, \ldots, M^j_{j-1}, a_j, M^j_{j+1}, \ldots, M^j_n)$".

[8] Actually, if $n \geq 3$, the other players may be able to keep player j's payoff even lower by using a correlated strategy against j, where the outcome of the correlating device is not observed by j (another way of putting this is to observe that, for $n \geq 3$, the inequality $\max_{a_j} \min_{a_{-j}} g_j(a_j, a_{-j}) \leq \min_{a_{-j}} \max_{a_j} g_j(a_j, a_{-j})$ can hold strictly). In keeping with the rest of the literature on repeated games, however, we shall rule out such correlated strategies.

Henceforth we shall normalize the payoffs of the game g so that $(v_1^*, \ldots, v_n^*) = (0, \ldots, 0)$. Let

$$U = \{(v_1, \ldots, v_n) | \exists (a_1, \ldots, a_n) \in A_1 \times \cdots \times A_n \text{ with } g(a_1, \ldots, a_n) = (v_1, \ldots, v_n)\},$$

$$V = \text{Convex Hull of } U,$$

and

$$V^* = \{(v_1, \ldots, v_n) \in V | v_i > 0 \text{ for all } i\}.$$

The set V consists of feasible payoffs, and V^* consists of feasible payoffs that Pareto dominate the minimax point. That is, V^* is the set of individually rational payoffs. In a repeated version of g, we suppose that players maximize the discounted sum of single period payoffs. That is, if $(a_1(t), \ldots, a_n(t))$ is the vector of actions played in period t and δ is player i's discount factor, then his payoff is $\sum_{t=1}^{\infty} \delta^{t-1} g_i(a_1(t), \ldots, a_n(t))$ and his average payoff is $(1-\delta) \sum_{t=1}^{\infty} \delta^{t-1} g_i(a_1(t), \ldots, a_n(t))$. We can now state a version of the Folk Theorem (see Hart **[10]** for more details).

THEOREM A (The Folk Theorem): *For any $(v_1, \ldots, v_n) \in V^*$, if players discount the future sufficiently little, there exists a Nash equilibrium of the infinitely repeated game where, for all i, player i's average payoff is v_i.*[9]

PROOF: Let $(s_1, \ldots, s_n) \in A_1 \times \cdots \times A_n$ be a vector of strategies[10] such that $g(s_1, \ldots, s_n) = (v_1, \ldots, v_n)$. Suppose that in the repeated game each player i plays s_i until some player j deviates from s_j (if more than one player deviates simultaneously, we can suppose that the deviations are ignored). Thereafter, assume that he plays M_i^j. These strategies form a Nash equilibrium of the repeated game if there is not too much discounting; any momentary gain that may accrue to player j if he deviates from s_j is swamped by the prospect of being minimaxed forever after. *Q.E.D.*

REMARK: If we disallowed correlated strategies, the same proof would establish that any positive vector in U could be enforced as an equilibrium. For other points $(v_1, \ldots, v_n)$ in V^* the statement of the theorem must be modified to read: for all $\varepsilon > 0$ there exists $\underline{\delta} < 1$ such that, for all $\delta > \underline{\delta}$, there exists a subgame perfect equilibrium of the infinitely repeated game in which each player i's average payoff is within ε of v_i, when players have discount factor δ. The ε qualification is needed because discounting and the requirement that each vector of actions be played an integral number of times limit the accuracy of approximating a correlated strategy by switching among action vectors over time.

[9] The hypothesis that the v_i's are positive is important, as a recent example by Forges, Mertens, and Neyman demonstrates.

[10] Or, if necessary, correlated strategies.

Of course, the strategies of Theorem A do not, in general, form a (subgame) perfect equilibrium (such an equilibrium is a configuration of strategies that form a Nash equilibrium in all subgames), because, if a player deviates, it may not be in others' interest to go through with the punishment of minimaxing him forever. However, Aumann and Shapley [3] and Rubinstein [20] showed that, when there is no discounting, the counterpart of Theorem A holds for perfect equilibrium.

THEOREM B (Aumann–Shapley/Rubinstein): *For any* $(v_1, \ldots, v_n) \in V^*$ *there exists a perfect equilibrium in the infinitely repeated game with no discounting, where, for all* i, *player* i*'s expected payoff each period is* v_i.[11]

REMARK: The Aumann-Shapley and Rubinstein arguments assume that past mixed strategies are observable (or, alternatively, that only pure strategies are ever played, which, in general, implies a smaller equilibrium set). However, the methods of Section 6 can be used to establish Theorem B in the case where only past actions are observable.

The idea of the proof is simple to express. Once again, as long as everyone has previously conformed, players continue to play their s_i's, leading to payoff v_i. If some player j deviates, he is, as before, minimaxed but, rather than forever, only long enough to wipe out any possible gain that he obtained from this deviation. After this punishment, the players go back to their s_i's. To induce the punishers to go through with their minimaxing, they are threatened with the prospect that, if any one of them deviates from his punishment strategy, he in turn will be minimaxed by the others long enough to make such a deviation not worthwhile. Moreover, *his* punishers will be punished if any one of *them* deviates, etc. Thus, there is a potential sequence of successively higher order punishments, where the punishment at each level is carried out for fear the punishment at the next level will be invoked.

Theorem B is not an exact counterpart of Theorem A because it allows no discounting at all (we investigate in Section 3 when an exact counterpart holds). Moreover, the strategies of the proof are a good deal more complex than those of Theorem A. One well-known case that admits both discounting and simple strategies is where the point to be sustained Pareto dominates the payoffs of a Nash equilibrium of the constituent game g.

THEOREM C (Friedman [5] and [6]): *If* $(v_1, \ldots, v_n) \in V^*$ *Pareto dominates the payoffs* $(y_1, \ldots, y_n)$ *of a (one-shot) Nash equilibrium* $(e_1, \ldots, e_n)$ *of* g, *then, if players discount the future sufficiently little, there exists a perfect equilibrium of the infinitely repeated game where, for all* i, *player* i*'s average payoff is* v_i.

[11] If there is no discounting, the sum of single-period payoffs cannot serve as a player's repeated game payoff since the sum may not be defined. Aumann and Shapley use (the lim infinum of) the average payoff; Rubinstein considers both this and the overtaking criterion, and the sketch of the proof we offer corresponds to this latter rule. (See also Hart [10]). The average payoff criterion allows more outcomes to be supported as equilibria than the overtaking criterion because for a player to strictly prefer to deviate he must gain in infinitely many periods. Indeed, for the former criterion, Theorem B holds for the closure of V^*.

PROOF: Suppose that players play actions that sustain $(v_1, \ldots, v_n)$ until someone deviates, after which they play $(e_1, \ldots, e_n)$ forever. With sufficiently little discounting, this behavior constitutes a perfect equilibrium. *Q.E.D.*

Because the punishments used in Theorem C are less severe than those in Theorems A and B, its conclusion is correspondingly weaker. For example, Theorem C does not allow us to conclude that a Stackelberg outcome can be supported as an equilibrium in an infinitely-repeated quantity-setting duopoly.

3. THE FOLK THEOREM IN INFINITELY REPEATED GAMES WITH DISCOUNTING

We now turn to the question of whether Theorem A holds for perfect rather than Nash equilibrium. Technically speaking, we are investigating the *lower hemicontinuity*[12] of the perfect equilibrium average payoff correspondence (where the independent variable is the discount factor, δ) at $\delta = 1$. We first remind the reader that this correspondence is *upper* hemicontinuous.[13]

THEOREM D: *Let $V(\delta) = \{(v_1, \ldots, v_n) \in V^* | (v_1, \ldots, v_n)$ are the average payoffs of a perfect equilibrium of the infinitely repeated game where players have discount factor $\delta\}$. The correspondence $V(\cdot)$ is upper hemicontinuous at any $\delta < 1$.*

It is easy to give examples where $V(\cdot)$ fails to be lower hemicontinuous at $\delta < 1$.

EXAMPLE 1: Consider the following version of the Prisoner's Dilemma:

	C	D
C	1, 1	−1, 2
D	2, −1	0, 0

For $\delta < 1/2$ there are no equilibria of the repeated game other than players' choosing D every period. However at $\delta = 1/2$ many additional equilibria appear, including playing C each period until someone deviates and thereafter playing D. Thus $V(\cdot)$ is not lower hemicontinuous at $\delta = 1/2$.

3A. *Two-Player Games*

Our particular concern, however, is the issue of lower hemicontinuity at $\delta = 1$, and we begin with two-player games. It turns out that, in this case, the exact analog of Theorem A holds for perfect equilibrium. We should point out, however,

[12] A correspondence $f: X \to Y$ is lower hemicontinuous at $x = \bar{x}$ if for any $\bar{y} \in f(\bar{x})$ and any sequence $x^m \to \bar{x}$ there exists a sequence $y^m \to \bar{y}$ such that $y^m \in f(x^m)$ for all m.

[13] If Y is compact the correspondence $f: X \to Y$ is upper hemicontinuous at $\bar{x}$ if for any sequence $x^m \to \bar{x}$ and any sequence $y^m \to \bar{y}$ such that $y^m \in f(x^m)$ for all m, we have $\bar{y} \in f(\bar{x})$.

that to establish this analog we *cannot* use the Aumann-Shapley/Rubinstein (AS/R) strategies of Theorem B once there is discounting. To see that, if there is discounting, such strategies may not be able to sustain all individually rational points, consider the following example.

EXAMPLE 2:

	C	D
C	1, 1	0, −2
D	−2, 0	−1, −1

For this game the minimax point is (0, 0), and so a "folk theorem" would require that we be able to sustain, in particular, strategies that choose (C, C) a fraction $(\varepsilon+1)/2$ of the time and (D, D) the remainder of the time, with $0<\varepsilon<1$ (note that for δ near 1, these strategies yield average payoffs of approximately $(\varepsilon, \varepsilon)$, which are individually rational). However such behavior cannot be part of an AS/R type of equilibrium.[14] Suppose, for example, that one of the players (say, player I) played C in a period where he was supposed to play D. In an AS/R equilibrium, player II would "punish" I by playing D sufficiently long to make I's deviation unprofitable. I's immediate gain from deviation is 1, and I's best response to D is C, resulting in a payoff 0. Therefore if the punishment lasts for t_1 periods, t_1 must satisfy

$$\frac{\delta\varepsilon(1-\delta^{t_1})}{1-\delta} > 1+t_1\cdot 0 = 1.$$

That is,

$$(1)\qquad t_1 > \frac{\log\left(\dfrac{\varepsilon\delta-1+\delta}{\delta\varepsilon}\right)}{\log\delta}.$$

Condition (1) can be satisfied as long as

$$(2)\qquad \delta > \frac{1}{1+\varepsilon}.$$

[14] Although Rubinstein's [19] theorem applies to stationary strategies without public correlation (and so does not directly imply that $(\varepsilon, \varepsilon)$ can be enforced without discounting) it does permit a continuum of strategies. Thus we can think of an "enlarged" version of this game, with a continuum of strategies indexed by ε, $0<\varepsilon<1$, as well as the original strategies C and D. Suppose that if both players play ε, they each receive ε. If one plays ε and the other plays ε', they each get 1/2. If player 2 plays ε and 1 plays C, 1 receives $1+\varepsilon$ and 2 receives 0. If 2 plays ε and 1 plays D, 1 gets $-1+\varepsilon$ and 2 gets ε. The payoffs are permuted if the roles are reversed. In this enlarged game, D is still the minimax strategy and (0, 0) is the minimax point. Hence, the pure strategy pair $(\varepsilon, \varepsilon)$ is sustainable by AS/R punishments without discounting, but, as we show in the text, these punishments fail in the discounting case.

But in order to punish player I, II must himself suffer a payoff of -2 for t_1 periods. To induce him to submit to such self-laceration, he must be threatened with a t_2-period punishment, where

$$-2\frac{(1-\delta^{t_1})}{1-\delta}+\frac{\delta^{t_1}\varepsilon(1-\delta^{t_2-t_1+1})}{1-\delta}>1.$$

That is,

$$t_2>-1+\log\frac{\delta^{t_1}\varepsilon-3+2\delta^{t_1}+\delta}{\varepsilon}\Big/\log\delta. \tag{3}$$

Such a t_2 exists as long as

$$\delta^{t_1}\varepsilon-3+2\delta^{t_1}+\delta>0,$$

which requires that

$$\delta>\left(\frac{2}{2+\varepsilon}\right)^{1/t_1}. \tag{4}$$

But (4) is a more stringent requirement than (2), since

$$(2/(2+\varepsilon))^{1/t_1}>\frac{1}{1+\varepsilon}.$$

Continuing iteratively, we find that, for successively higher order punishments, δ is bounded below by a sequence of numbers converging to 1. Since δ is itself strictly less than 1, however, this is an impossibility, and so an AS/R equilibrium is impossible.

The problem is that in this example the punisher is hurt more severely by his punishment than is his victim. He must therefore be threatened with an even stronger punishment. Without discounting, this can be arranged by (roughly) taking the t_i's to be a geometric series, as in Rubinstein [20]. With discounting, however, arbitrarily long punishments are not arbitrarily severe, because far-off punishments are relatively unimportant.

These punishment strategies are not "simple" in the sense of Abreu [1] because they are not independent of history, i.e., they depend on the previous sequence of deviations. Abreu's work shows that there is no loss in restricting attention to simple punishments when players discount the future. Indeed, we make use of simple punishments in the proof of the following result, which shows that we can do without arbitrarily severe punishments in the two-player case.

THEOREM 1: *For any* $(v_1, v_2)\in V^*$ *there exists* $\underline{\delta}\in(0,1)$ *such that, for all* $\delta\in(\underline{\delta},1)$, *there exists a subgame perfect equilibrium of the infinitely repeated game in which player i's average payoff is* v_i *when players have discount factor* δ.

PROOF: Let M_1 be player one's minimax strategy against two, and M_2 a minimax strategy against one. Take $\bar{v}_i=\max_{a_1,a_2} g_i(a_1,a_2)$. For $(v_1, v_2)\in V^*$ choose $\underline{v}$ and $\underline{\delta}$ such that for $i=1,2$,

$$v_i>\bar{v}_i(1-\underline{\delta})+\underline{\delta}\underline{v}_i^{**}, \tag{5}$$

where

$$v_i^{**} = (1-\underline{\delta}^{\underline{\nu}})g_i(M_1, M_2) + \underline{\delta}^{\underline{\nu}} v_i, \tag{6}$$

with

$$v_i^{**} > 0. \tag{7}$$

To see that $\underline{\nu}$ and $\underline{\delta}$ exist, choose $\underline{\delta}$ close enough to 1 so that

$$v_i > \bar{v}_i(1-\underline{\delta}) \tag{5a}$$

and (7) holds when $\underline{\nu} = 1$. If with $\underline{\nu} = 1$, (5) is violated, consider raising $\underline{\nu}$. From (5a), (5) will eventually be satisfied. Since, for $\underline{\delta}$ close to 1, (6) declines almost continuously as $\underline{\nu}$ increases, by taking $\underline{\delta}$ near enough to one we can ensure that (7) will be satisfied for the first $\underline{\nu}$ for which (5) holds.

Condition (5) guarantees that player i prefers receiving v_i forever to receiving his maximum possible payoff ($\bar{v}_i$) once, then receiving $g_i(M_1, M_2)$ for $\underline{\nu}$ periods, and receiving v_i thereafter. Condition (7) ensures that being punished for deviating is still better than receiving the reservation value, zero, forever. Clearly, for any $\delta > \underline{\delta}$ there is a corresponding $\nu(\delta)$ such that (5) and (7) hold for $(\delta, \nu(\delta))$.

Let (s_1, s_2) be correlated one-shot strategies corresponding to (v_1, v_2): $g_i(s_1, s_2) = v_i$. Consider the following repeated game strategies for player i:

(A) Play s_i each period as long as (s_1, s_2) was played last period.

After any deviation from (A):

(B) Play $M_i\nu(\delta)$ times and then start again with (A). If there are any deviations while in phase (B), then begin phase (B) again.

These strategies form a subgame-perfect equilibrium. Condition (5) guarantees that deviation is not profitable in phase (A). In phase (B), player i receives an average payoff of at least v_i^{**} by not deviating. If he deviates, he can obtain at most 0 the first period (because his opponent, j, is playing M_j), and thereafter can average at most v_i^{**}. Hence deviation is not profitable in phase (B). *Q.E.D.*

The statement of Theorem 1 assumes that correlated strategies are available. To see how the theorem must be modified if they are not, see the remark following Theorem A.

The strategies in the proof of Theorem 1 are easily summarized. After a deviation by either player, each player minimaxes the other for a certain number of periods, after which they return to the original path. If a further deviation occurs during the punishment phase, the phase is begun again.

Notice that in the proof of Theorem 1 the only place where we invoked our assumption that past mixed strategies can be observed was in supposing that deviations from the minimax strategies, M_1 and M_2, can be detected. This assumption is dropped in Section 6.

3B. *Three or More Players*

The method we used to establish Theorem 1—"mutual minimaxing"—does not extend to three or more players. This is because with, say, three players there

may exist no triple of alternatives (M_1, M_2, M_3) such that M_2 and M_3 minimax player one, M_1 and M_3 minimax two, and M_1 and M_2 minimax three; that is, the "mutual minimax" property may fail. However, the situation is even worse: not only does the method of proving Theorem 1 fail to extend, but the result itself does not generalize. To see this, consider the following example.

EXAMPLE 3:

1, 1, 1	0, 0, 0
0, 0, 0	0, 0, 0

0, 0, 0	0, 0, 0
0, 0, 0	1, 1, 1

In this game, player one chooses rows, player two chooses columns, and three, matrices. Note that whatever one player gets, the others get too.

CLAIM: For any $\delta < 1$ there does *not* exist a perfect equilibrium of the supergame in which the average payoff ε is less than 1/4 (the one-shot mixed strategy equilibrium payoff).

PROOF: For fixed $\delta < 1$, let $\alpha = \inf\{\varepsilon | \varepsilon$ sustainable as an average payoff of a player in a perfect equilibrium$\}$. We must show that $\alpha \geq 1/4$. Let

$$\beta = \min_{\sigma_1^*, \sigma_2^*, \sigma_3^*} \max_{\sigma_1, \sigma_2, \sigma_3} \{g_1(\sigma_1, \sigma_2^*, \sigma_3^*), g_2(\sigma_1^*, \sigma_2, \sigma_3^*),$$

$$g_3(\sigma_1^*, \sigma_2^*, \sigma_3)\}.$$

That is, β is the minimum that the most fortunate defector can obtain in an optimal (one-shot) deviation from an arbitrary configuration of strategies. We claim that $\beta \geq 1/4$. Hence, the mutual minimax property does not hold.

To see this, let a_i be the probability that player i plays the "first" pure strategy, i.e., the first column, row, or matrix as appropriate. For some player i it must be the case that, for $j \neq i \neq k$, either $a_j \geq 1/2$ and $a_k \geq 1/2$ or $a_j \leq 1/2$ and $a_k \leq 1/2$. But since player i can obtain any convex combination of $a_j a_k$ and $(1-a_j)(1-a_k)$ as a payoff, he can get a payoff of at least 1/4.

Thus in any equilibrium, the payoff to deviating for some player is at least $1/4 + \delta\alpha/(1-\delta)$. Let $\{\varepsilon_m\}$ be a sequence of possible average payoffs in perfect equilibria, where $\varepsilon_m \to \alpha$. For all m we have $1/4 + \delta\alpha/(1-\delta) \leq \varepsilon_m/(1-\delta)$. Hence, $1/4 + \delta\alpha/(1-\delta) \leq \alpha/(1-\delta)$, and so $\alpha \geq 1/4$. *Q.E.D.*

The game of Example 3 is degenerate in the sense that V^*, the individually rational set, is one-dimensional. This degeneracy is responsible for the discontinuity in $V(\delta)$ as the next result demonstrates.

THEOREM 2: *Assume that the dimensionality of V^* equals n, the number of players, i.e., that the interior of V (relative to n-dimensional space) is nonempty. Then, for any $(v_1, \ldots, v_n)$ in V^*, there exists $\underline{\delta} \in (0, 1)$ such that for all $\delta \in (\underline{\delta}, 1)$ there exists a subgame-perfect equilibrium of the infinitely repeated game with discount factor δ in which player i's average payoff is v_i.*

The idea behind the proof of Theorem 2 is simple. If a player deviates, he is minimaxed by the other players long enough to wipe out any gain from his deviation. To induce the other players to go through with minimaxing him, they are ultimately given a "reward" in the form of an additional "ε" in their average payoff. The possibility of providing such a reward relies on the full dimensionality of the payoff set.

PROOF: Choose $s = (s_1, \ldots, s_n)$ so that $g(s_1, \ldots, s_n) = (v_1, \ldots, v_n)$ (again we allow correlated strategies). Also choose $(v_1', \ldots, v_n')$ in the interior of V^* such that $v_i > v_i'$ for all i. Since $(v_1', \ldots, v_n')$ is in the interior of V^* and V^* has full dimension, there exists $\varepsilon > 0$ so that, for each j,

$$(v_1' + \varepsilon, \ldots, v_{j-1}' + \varepsilon, v_j', v_{j+1}' + \varepsilon, \ldots, v_n' + \varepsilon)$$

is in V^*. Let $T^j = (T_1^j, \ldots, T_n^j)$ be a joint strategy that realizes these payoffs. Let $M^j = (M_1^j, \ldots, M_n^j)$ be an n-tuple of strategies such that the strategies for players other than j together minimize player j's maximum payoff, and such that $g_j(M^j) = 0$. Let $w_i^j = g_i(M^j)$ be player i's per-period payoff when minimaxing player j. For each i choose an integer ν_i such that

$$\frac{\bar{v}_i}{v_i'} < 1 + \nu_i, \tag{8}$$

where, as before, $\bar{v}_i$ is player i's greatest one-shot payoff.

Consider the following repeated game strategy for player i:

(A) play s_i each period as long as s was played last period. If player j deviates from (A),[15] then:

(B) play M_i^j for ν_j periods, and then:

(C) play T_i^j thereafter.

If player k deviates in phase (B) or (C), then begin phase (B) again with $j = k$.[16]

If player i deviates in phase (A) and then conforms, he receives at most $\bar{v}_i$ the period he deviates, zero for ν_i periods, and v_i' each period thereafter. His total payoff, therefore, is no greater than

$$\bar{v}_i + \frac{\delta^{\nu_i + 1}}{1 - \delta} v_i'. \tag{9}$$

[15] If several players deviate from (A) simultaneously, then we can just as well suppose that everyone ignores the deviation and continues to play s.

[16] As in footnote 15, we can suppose that simultaneous deviation by several players is ignored.

If he conforms throughout, he obtains $v_i/(1-\delta)$, so that the gain to deviating is less than

$$(10) \qquad \bar{v}_i - \frac{1-\delta^{(\nu_i+1)}}{1-\delta} v_i^l.$$

Because $(1-\delta^{\nu_i+1})/(1-\delta)$ converges to ν_i+1 as δ tends to 1, condition (8) ensures that (10) is negative for all δ larger than some $\underline{\delta} < 1$. If player i deviates in phase (B) when *he* is being punished, he obtains at most zero the period in which he deviates, and then only lengthens his punishment, postponing the positive payoff v_i^l. If player i deviates in phase (B) when player j is being punished, and then conforms, he receives at most

$$\bar{v}_i + \frac{\delta^{\nu_j+1}}{(1-\delta)} v_i^l,$$

which is less than $\bar{v}_i + v_i^l/(1-\delta)$. If, however, he does not deviate, he receives at least

$$w_i^j \frac{(1-\delta^{\nu})}{1-\delta} + \frac{\delta^{\nu+1}}{1-\delta}(v_i^l + \varepsilon),$$

for some ν between 1 and ν_j. Thus the gain to deviating is at most

$$(11) \qquad \bar{v}_i + \frac{(1-\delta^{\nu+1})}{1-\delta}(v_i^l - w_i^j) - \frac{\delta^{\nu+1}}{1-\delta}\varepsilon - \delta^{\nu} w_i^j.$$

As $\delta \to 1$, the second term in (11) remains finite because $(1-\delta^{\nu+1})/(1-\delta)$ converges to $\nu+1$. But, because δ^{ν} converges to 1, the third converges to negative infinity. Thus there exists $\underline{\delta}_i < 1$ such that for all $\delta > \underline{\delta}_i$, player i will not deviate in phase (B) if the discount factor is δ.

Finally, the argument for why players do not deviate in phase (C) is practically the same as that for phase (A). *Q.E.D.*

4. INCOMPLETE INFORMATION WITH NASH THREATS

Suppose that a game is repeated finitely many times, ν, that players maximize the (expected) sum of their one-shot payoffs, and that players can observe all past one-shot strategies (including mixed strategies). This repeated game can be embedded in a ν-period sequential game of incomplete information. Suppose that players' payoffs and, perhaps, even their action spaces A_i depend on their types (although we shall not explicitly consider this latter type of incomplete information). With probability, say, $1-\varepsilon$, a given player i is described by g_i. We call a player of this type "sane" or "rational." However with probability ε his payoffs and action spaces may be different and might even be period-dependent. Such a player we call "crazy." The motivation for suggesting this possibility is that often one cannot be sure what kind of player one is up against. One might

be *almost* sure, but even if ε is nearly zero, one may nevertheless wish to take into account other possibilities. Indeed, as the following result shows any vector of payoffs Pareto dominating a Nash equilibrium of the constituent game, g, can arise approximately[17] as the average payoffs of a perfect equilibrium of a game of incomplete information[18] that, with high probability is just a finitely repeated version of g. The result, therefore, is the counterpart for finitely repeated games of incomplete information of Friedman's Theorem C above.

THEOREM 3: *Let* $(e_1, \ldots, e_n)$ *be a Nash equilibrium of the game,* g, *and let* $(y_1, \ldots, y_n) = g(e_1, \ldots, e_n)$. *For any* $\varepsilon > 0$ *and any* $(v_1, \ldots, v_n) \in V^*$ *such that* $v_i > y_i$ *for all* i, *there exists* $\underline{\nu}$ *such that for any* $\nu > \underline{\nu}$ *there exists a* ν*-period sequential game where, with probability* $1-\varepsilon$, *player* i *is described in each period by* g_i *and in which there exists a sequential equilibrium where player* i*'s average payoff is within* ε *of* v_i.

REMARK: Notice that the theorem asserts the existence of a *game* as well as of an equilibrium. This enables us to choose the form of the incomplete information.

PROOF: As above, let $\bar{v}_i = \max_{a_1,\ldots,a_n} g_i(a_1, \ldots, a_n)$. Also define $\underline{v}_i = \min_{a_1,\ldots,a_n} g_i(a_1, \ldots, a_n)$. Choose $s = (s_1, \ldots, s_n)$ so that $g(s_1, \ldots, s_n) = (v_1, \ldots, v_n)$.

We will consider a sequential game where each player i can be of two types: "sane," in which case his payoffs are described by g_i, and "crazy," in which case he plays s_i each period as long as s has always been played previously and otherwise plays e_i. Players initially attach probability ε to player i's being crazy and probability $1-\varepsilon$ to i's being sane. We shall see that early enough in the game, both types of player i play s_i if there have been no deviations from s. Hence, a deviation from s_i constitutes an "impossible" event, one for which we cannot apply Bayes' rule, and so we must specify players' beliefs about i in such an event. We shall suppose that then all players attach probability one to player i's being sane.

Now starting at any point of this sequential game where there has already been a deviation from s, it is clear that one sequential equilibrium of the continuation game consists of all players playing Nash strategies (the e_i's) until the end of the game. We shall always select this equilibrium.

[17] The qualification "approximately" is necessary because the game is repeated only finitely more times.

[18] Because the game is one of incomplete information, we must use some sort of Bayesian perfect equilibrium concept. We shall adopt the sequential equilibrium of Kreps and Wilson [15]. According to this concept a player has probabilistic beliefs about other players' types that are updated in Bayesian fashion according to what other players do. An equilibrium is a configuration of strategies as functions of players' types such that, at every point of the game, each player's strategy is optimal for him, given others strategies and his beliefs about their types (actually the concept is a bit more refined than this, but, given the simple structure of our games, this description will do).

Choose $\underline{\nu}$ so that

$$(12) \qquad \underline{\nu} > \max_i \left[\frac{\bar{v}_i - (1-\varepsilon^{n-1})\underline{v}_i}{\varepsilon^{n-1}(v_i - y_i)} \right].$$

We will show that in a period with ν periods remaining in the game, where $\nu \geq \underline{\nu}$, a sane player of type i will play s_i if there have been no deviations from s to that point. If that period he plays something other than s_i, his maximum payoff is $\bar{v}_i$. Subsequently his payoff is y_i every period, since, starting from any point after a deviation from s, we always select the "Nash" sequential equilibrium. Thus, if he deviates from s_i with ν periods remaining, an upper bound to i's payoff for the rest of the game is

$$(13) \qquad \bar{v}_i + (\nu - 1)y_i.$$

Suppose, on the other hand, he uses the sequential strategy of playing s_i each period until someone deviates from s and thereafter playing e_i. In that case, his payoff is v_i each period for the rest of the game if the other players are all crazy. If at least one of the other players is not crazy, the worst that could happen to i is that his payoff is $\underline{v}_i$ in the first period and y_i in each subsequent period. Now, assuming that there have been no previous deviations from s, the probability that all the others are crazy is ε^{n-1}. Hence, a lower bound to i's payoff if he uses this sequential strategy is

$$(14) \qquad \varepsilon^{n-1}\nu v_i + (1-\varepsilon^{n-1})(\underline{v}_i + (\nu - 1)y_i).$$

From (12), (14) is bigger than (13). Hence all players i will play s_i in any period at least $\underline{\nu}$ periods from the end. Thus, for any $\varepsilon > 0$, we can choose ν big enough so that player i's average payoff of the ν-period sequential game is within ε of v_i. *Q.E.D.*

5. THE FOLK THEOREM IN FINITELY REPEATED GAMES OF INCOMPLETE INFORMATION

In this section we strengthen the result of Section 4 by showing roughly that any individually rational point can be sustained (approximately) as the average equilibrium payoffs of a finitely repeated game if the number of repetitions is large enough. This assertion is not quite true for the same reason that the perfect equilibrium counterpart to Theorem A does not hold for three or more players: a discontinuity in $V(\delta)$ can occur if the payoff set is degenerate. For this reason we confine attention to two-player games.[19]

THEOREM 4: *For any* $(v_1, v_2) \in V^*$ *and any* $\varepsilon > 0$ *there exists* $\underline{\nu}$ *such that for any* $\nu > \underline{\nu}$ *there exists a* ν*-period sequential game such that, with probability* $1-\varepsilon$,

[19] If we posited full dimension we could also establish the result for three or more players; i.e., we could establish the analog of Theorem 3.

player i is described in each period by g_i and there exists a sequential equilibrium where player i's average payoff is within ε of v_i.

The proof we provide in this section assumes the existence of a one-shot Nash equilibrium that yields both players strictly more than their minimax values. We have established the theorem in general using a similar but more complex argument that is presented in our 1985 working paper.

Briefly, the proof goes as follows: we know that with an infinite horizon our "mutual minimax" strategies of Theorem 1 will enforce any individually rational outcome. The problem with a finite horizon is to avoid the familiar "backwards unraveling" of these strategies from the end. To do so, we introduce the probability ε that a player is "crazy" and will punish his opponent for deviations that would otherwise be too near the end to be deterred by credible (i.e., sequentially rational) threats. More specifically, we partition the game into three "phases." In the first phase, Phase I, players follow the strategies of Theorem 1. That is, they play strategies enforcing the desired outcome unless someone deviates, which triggers mutual minimaxing for β periods, followed by a return to the original path. Deviations during the punishment period restart the mutual punishment. Phase II is a transitional phase. Punishments begun in Phase I are continued, if necessary, in Phase II, but deviations in Phase II are ignored until Phase III. In Phase III, a crazy type plays a Nash equilibrium strategy unless his opponent deviated in Phase II, in which case he plays his minimax strategy. Phase III is an "endgame" in which the crazy types create punishments that do not unravel, and Phase II simply connects this endgame to the strategies of Phase I. The proof shows that by making the last two phases long enough we indeed have an equilibrium, and, moreover, that the required lengths are independent of the total length of the game. Thus if the game lasts long enough, Phase I constitutes most of the game, and our result follows.[20]

PROOF: Let $x_i = g_i(M_1, M_2)$. Clearly $x_i \leq 0$. Let (y_1, y_2) be the expected payoffs to a Nash equilibrium (e_1, e_2) of the one-shot game g, and assume y_1 and y_2 are strictly positive.

As before, we suppose that players can use correlated mixed strategies. Let (s_1, s_2) be correlated strategies yielding payoffs (v_1, v_2). Let β be an integer such that

$$\beta \geq \max_i (\bar{v}_i / v_i), \tag{15}$$

and, as before, let $\underline{v}_i = \min_{a_1, a_2} g_i(a_1, a_2)$. For given $\varepsilon > 0$, choose an integer α_i so that

$$\beta \bar{v}_i + \alpha_i(1-\varepsilon) y_i < \alpha_i y_i + \underline{v}_i + \beta x_i \tag{16}$$

and take $\alpha = \max_i \alpha_i$.

[20] We thank a referee for suggesting this simplified form of our earlier proof.

To describe the equilibrium play and the "crazy" player types, we partition the game into three "phases." We will number the periods so that the game ends in period 1. Phase I runs from period ν to period $\alpha+\beta+1$, Phase II from $(\alpha+\beta)$ to $\alpha+1$, and Phase III from α to 1.

We will specify crazy behavior recursively, that is, in each period we specify how the crazy player will behave if play to that point has corresponded to the "crazy" play specified for the previous periods, and also how the crazy player will respond to any deviation from that behavior.

Let us begin with Phase I. We define the index $\Psi(t)$ as follows. Set $\Psi(\nu)=\nu+\beta$. In period t, $\nu \geq t > \alpha+\beta$, the crazy type (of player i) plays s_i if $\Psi(t)-t \geq \beta$, and M_i otherwise. We set $\Psi(t)=\Psi(t+1)$ if there was no deviation from "crazy" behavior in period $t+1$, and $\Psi(t)=t$ otherwise. Thus the crazy type plays s_i until someone deviates. Deviations trigger β periods of minimaxing followed by a return to s_i if there have been no further deviations. Any deviation restarts the mutual punishment portion of the sequence, which runs for β periods after the deviation.

The crazy type follows the same strategy in Phase II as in Phase I, except that deviations in this phase do not change the index $\Psi(t)$. More specifically, in Phase II, $\Psi(t)=\Psi(t+1)$ regardless of play in period $t+1$, and the crazy type plays s_i if $\Psi(t)-t \geq \beta$, and plays M_i otherwise. Deviations in Phase II influence behavior in Phase III through a second index variable, Θ. This index has four possible values: $\Theta=0$ if there have been no deviations from crazy behavior in Phase II; $\Theta=1$ if only player one deviated; $\Theta=2$ if only player two deviated; and $\Theta=b$ if both players have deviated. The index Θ is not changed by deviations in Phase III. In Phase III the crazy type plays e_i if $\Theta=0$, i, or b, and plays M_i if $\Theta=j$. That is, the crazy type of player i punishes his opponent in Phase III for having deviated in Phase II *unless* player i himself also deviated.

Next we describe the behavior of the "sane" types of each player. For a sequential equilibrium we must specify both a strategy for each player, mapping observations into actions, and a system of beliefs, mapping observations into inferences. In Phase I, each sane type's strategy is the same as the corresponding crazy strategy. If his opponent deviates from crazy behavior, the sane player's beliefs are unchanged—he continues to assign the *ex ante* probabilities of ε and $1-\varepsilon$, respectively, to his opponent being crazy or sane.

In Phase II, if, in state $\Theta=0$, a player deviates from crazy behavior, his opponent attaches probability one to his being crazy. The strategy of the sane type (of player i) in Phase II if $\Theta=0$ or j is to play as a crazy type. We do not specify sane play if $\Theta=i$ or b.

In Phase III, the sane type plays e_i if $\Theta=0$ or j. If player i did not deviate in Phase II, then his beliefs are not changed by play in Phase III. If player i did deviate in Phase II, and player j plays M_j at the beginning of Phase III, j is revealed to be crazy, while if j plays e_j, j is revealed to be sane. We do not specify sane behavior for Phase III if $\Theta=i$ or b except to require that it depend on past outcomes only through the player's beliefs and Θ. Thus we choose some equilibrium for each set of initial beliefs and Θ. The exact nature of this behavior and

TABLE I

Phase	Θ	State $\Psi(t)$	Strategies crazy	Strategies sane	Beliefs (probability i attaches to j's being crazy)
I (periods v to $\alpha+\beta+1$)	—	$\geq\beta+t$	s_i	s_i	ε
	—	$<\beta+t$	M_i	M_i	ε
II (periods $\alpha+\beta$ to $\alpha+1$)	0	$\geq\beta+t$	s_i	s_i	ε
	0	$<\beta+t$	M_i	M_i	ε
	j	$\geq\beta+t$	s_i	s_i	1
	j	$<\beta+t$	M_i	M_i	1
	i	$\geq\beta+t$	s_i	?	?
	i	$<\beta+t$	M_i	?	?
	b	$\geq\beta+t$	s_i	?	?
	b	$<\beta+t$	M_i	?	?
III (periods α to 1)	0		e_i	e_i	ε
	j		M_i	e_i	1
	i		e_i	?	?
	b		e_i	?	?

the behavior in Phase II if $\Theta = i$ or b is irrelevant for our analysis. We know there must exist an equilibrium for each such subgame,[21] and, by deriving upper bounds on player i's payoffs there, we will show that these subgames are not reached on the equilibrium path. Thus, regardless of the form of this "endplay," there is a sequential equilibrium of the whole game in which sane types play as described in Phase I. The specified behavior and beliefs are summarized in Table I.

Now we must show that the specified strategies form a Nash equilibrium in each subgame, and that the beliefs in each period are consistent with Bayes rule. We shall consider whether player one's specified behavior is optimal given his beliefs and player two's specified behavior.

We begin in Phase III. If $\Theta = 0$ or 2, player one expects his opponent to play the Nash strategy e_2 for the duration of the game (recall that if $\Theta = 2$, player one believes player two is crazy), so that the best player one can do is to play his Nash strategy e_1.

Now consider some period t in Phase II, i.e., $\alpha+\beta \geq t > \alpha$. First assume $\Theta = 0$. If player one conforms to his specified strategy in Phase II, his payoff each period is either v_1 (if $\Psi(t)-t \geq \beta$) or x_1 (if $\Psi(t)-t < \beta$). Thus his lowest possible expected payoff for the remainder of Phase II is $(t-\alpha)x_1$. If he sticks to specified behavior in Phase III as well, he receives αy_1. Thus if player one conforms from period t in Phase II onwards he receives at least

$$(17) \qquad (t-\alpha)x_1 + \alpha y_1.$$

If however player one deviates in Phase II, his highest payoff in that phase is $(t-\alpha)\bar{v}_1$. Then in Phase III, player two plays M_2 if crazy, and e_2 if sane. Thus

[21] To establish this we can appeal to the existence theorem of Kreps–Wilson [15], since g is a finite game.

an upper bound to player one's expectation in Phase III is $(1-\varepsilon)\alpha y_1$, and the total payoff to deviating in period t of Phase II is at most

(18) $\qquad (t-\alpha)\bar{v}_1+(1-\varepsilon)\alpha y_1.$

Since t is in Phase II, $t-\alpha \leqslant \beta$, and the equation defining α, (16), ensures that deviation is unprofitable.

If $\Theta=2$ in Phase II, player one is sure that player two is crazy. Thus if player one follows his specified strategy, his payoff is again bounded by (17), while if he deviates in period t his payoff is at most

(19) $\qquad (t-\alpha)\bar{v}_1+\alpha\cdot 0.$

Once more, formula (16) ensures that α is large enough so that deviation is unprofitable.

Finally consider a period t in Phase I. From our specification, deviations in Phase I do not change the players' beliefs or the value of Θ. Thus from our previous analysis, both players will conform in Phases II and III regardless of the play in Phase I, so that any sequence of deviations must end at the start of Phase II.

First assume that $\Psi(t)<\beta+t$, so that t is part of a "punishment sequence." If player one conforms in period t and subsequently, his payoff is

(20) $\qquad (t-\Psi(t)+\beta)x_1+(\Psi(t)-\alpha-\beta)v_1+\alpha y_1.$

If player one deviates in period t and thereafter conforms, his maximum payoff in period t is zero, and he endures the "punishment" of x_1 for the next β periods, so his payoff is at most

(21) $\qquad \beta x_1+(t-\beta-\alpha-1)v_1+\alpha y_1,$

which is less than (20). In particular, player one would never deviate in the last period of Phase I, and, by backwards induction, will not wish to deviate in period t.

Last assume $\Psi(t)\geqslant\beta+t$, so that player two plays s_2 in period t. If player one deviates in period t but conforms thereafter, he receives at most

(22) $\qquad \bar{v}_1+\beta x_1+(t-\alpha-\beta-1)v_1+\alpha y_1.$

If player one conforms to his prescribed strategy, he receives

(23) $\qquad (t-\alpha)v_1+\alpha y_1.$

The gain to deviating, the difference between (22) and (23), is thus

(24) $\qquad \bar{v}_1+\beta x_1-(\beta+1)v_1.$

Since x_1 is nonpositive, formula (15) defining β ensures that (24) is negative, so player one will not deviate. Thus the specified strategies are indeed in equilibrium. This equilibrium will yield the payoff (v_1, v_2) for $v-\alpha-\beta$ periods, so that by taking ν sufficiently large we can make each player i's average payoff arbitrarily near v_i. *Q.E.D*

Notice that in the proof of Theorem 3 we not only chose the form of "crazy" behavior to suit our needs, but also selected particular conjectures for sane players

when Bayes' rule is inapplicable. We should emphasize that our choice of conjectures was not arbitrary; the theorem is not true if, for example, a player believes his opponent to be *sane* with probability one after a deviation.

Kreps [12], moreover, has pointed out that, because of our choice of conjectures, our equilibrium may not be stable in the sense of Kohlberg-Mertens [11].[22] In response, we offer the following modified version of our construction. This version has no zero probability events, so that the issue of the "reasonableness" of the conjectures and the stability of the equilibrium do not arise. Specifically, assume that at each period in Phase II a crazy player plays as before with probability $(1-\mu)$, while assigning strictly positive probability to every other pure strategy. If μ is sufficiently near zero, the expected payoffs in every subgame are essentially unchanged, and our strategies are still in equilibrium. Given that the crazy player "trembles" with positive probability in Phase II, any deviation in that phase must reveal that the deviator is crazy, as we specified.

6. UNOBSERVABLE MIXED STRATEGIES

The arguments in Sections 2–5 rely on mixed strategies' being observable. Although this assumption is often used, at least implicitly, in the Folk Theorem literature and can be justified in some circumstances, the more natural hypothesis is that only the moves that players actually make are observed by their opponents. In this section we argue that our results continue to hold with unobservable mixed strategies.

We suggested earlier that the only significant use that our proofs make of the assumption that mixed strategies are observable is in supposing that *minimax* strategies are observable. The heart of the argument, in Theorem 5, therefore, is to show that it suffices for other players to observe the realization of a punisher's random mixed strategy.

Although we rule out observation of private mixed strategies, we continue to assume, for convenience, that strategies can depend on the outcome of publicly observed random variables. We also impose the nondegeneracy assumption of Theorem 2.

THEOREM 5: *Theorem 2 continues to hold when we assume that players can observe only the past actions of other players rather than their mixed strategies.*

PROOF: Choose s, $(v_1^*, \ldots, v_n^*)$, $(v_1, \ldots, v_n)$, $(M^1, \ldots, M^n)$, and w_i^j, $i, j = 1, \ldots, n$ as in the proof of Theorem 2. For each i and j, consider M_i^j, player i's minimax strategy against j. This strategy is, in general a randomization among the m_i^j pure strategies $\{a_i^j(k)\}_{k=1}^{m_i^j}$, where we have chosen the indexation so that,

[22] The intuitive basis for Krep's observation is that since the crazy types prefer crazy play, the sane types are "more likely" to deviate from it. Of course, in the games as specified this is not strictly true, but in the "perturbed" versions of the game considered when testing for stability, there would be some deviations that did not increase the opponents's belief that the deviator is crazy.

for each $k = 1, \ldots, m_i^j - 1$,

$$g_i(a_i(k), M_{-i}^j) \leqslant g_i(a_i^j(k+1), M_{-i}^j).$$

For each k, let

$$p_i^j(k) = g_i(a_i^j(k), M_{-i}^j) - g_i(a_i^j(1), M_{-i}^j).$$

The repeated game strategies we shall consider closely resemble those in the proof of Theorem 2. Player i:

(A) plays s_i each period as long as s was played the previous period. If player j deviates from (A), then player i:

(B) plays M_i^j for ν_j periods.

If player i plays pure strategy $a_i^j(k)$ in periods $t_1, \ldots, t_m$ of phase (B), define

$$r_i^j(k) = \sum_{h=1}^{m} \delta^{t_h - 1} p_i^j(k).$$

Thus, $r_i^j(k)$ is the expected "bonus" that player i obtains from playing $a_i^j(k)$ rather than $a_i^j(1)$ in those periods. Take $r_i^j = \sum_k r_i^j(k)$. Then, r_i^j is the total expected bonus from phase (B). Let

$$z_i^j = \frac{r_i^j(1-\delta)}{\delta^{\nu_j}}.$$

Because $(v_1', \ldots, v_n')$ is the interior of V^* and V^* has full dimension, there exists $\varepsilon > 0$ so that, for each j,

$$(v_1' + \varepsilon, \ldots, v_{j-1}' + \varepsilon, v_j', v_{j+1}' + \varepsilon, \ldots, v_n' + \varepsilon)$$

is in V^*. Since z_i^j tends to zero as δ tends to 1, we can choose δ big enough so that, for all i and j, $z_i^j < \varepsilon/2$. Then

$$(25) \qquad (v_1' + \varepsilon - z_1^j, \ldots, v_{j-1}' + \varepsilon - z_{j-1}^j, v_j', v_{j+1}' + \varepsilon - z_{j+1}^j, \ldots, v_n' + \varepsilon - z_n^j)$$

is in V^*. If player h deviates from the prescribed behavior in phase (B) the phase is begun again with $j = h$. Player i cannot detect whether player h has deviated from M_h^j, but he can observe whether h has deviated from the *support* of M_h^j. Accordingly, if h so deviates, player i begins phase (B) again with $j = h$. Let $T^j(z) = (T_1^j(z), \ldots, T_n^j(z))$ be a vector of strategies that realizes the payoffs (25) (note that $T^j(z)$ depends on the particular realization of pure strategies in phase (B)). Now suppose that at the conclusion of phase (B), player i:

(C) Plays $T_i^j(z)$ thereafter, and, if player h deviates from (C), then i begins phase (B) again with $j = h$.

The strategies T^j are chosen so that player i will be indifferent among all the pure strategies in the support of M_i^j. The idea is that any expected advantage that player i obtains from using $a_i^j(k)$ rather than $a_i^j(i)$ in phase (B) is subsequently removed in phase (C). Player i then may as well randomize as prescribed by M_i^j. He will not deviate from the support of M_i^j since such a deviation will be detected and punished. *Q.E.D.*

We can also show that Theorem 1 continues to hold with unobservable mixed strategies; we omit the details, except to say that our proof relies on "rewarding" a player who uses a "costly" element of his minimax set with a (small) probability that play will switch from mutual minimaxing to a static Nash equilibrium.

University of California, Berkeley
and
Harvard University

REFERENCES

[1] ABREU, D.: "Repeated Games with Discounting," Ph.D. dissertation, Department of Economics, Princeton University, 1983.
[2] AUMANN, R.: "Subjectivity and Correlation in Randomized Strategies," *Journal of Mathematical Economics*, 1(1974), 67–96.
[3] AUMANN, R., AND L. SHAPLEY: "Long Term Competition: A Game Theoretic Analysis," mimeo, Hebrew University, 1976.
[4] BENOIT, J. P., AND V. KRISHA: "Finitely Repeated Games," *Econometrica*, 53(1985), 890–904.
[5] FRIEDMAN, J.: "A Noncooperative Equilibrium For Supergames," *Review of Economic Studies*, 38(1971), 1–12.
[6] ———: *Oligopoly and the Theory of Games*. Amsterdam: North-Holland, 1977.
[7] ———: "Trigger Strategy Equilibria in Finite Horizon Supergames," mimeo, University of North Carolina, Chapel Hill, 1984.
[8] FUDENBERG, D., AND E. MASKIN: "Nash and Perfect Equilibrium Payoffs in Discounted Repeated Games," mimeo, Harvard University, 1986.
[9] ———: "The Folk Theorem in Repeated Games with Discounting and with Incomplete Information," MIT Working Paper, 1985.
[10] HART, S.: "Lecture Notes: Special Topics in Game Theory," Stanford IMSSS Technical Report, 1979.
[11] KOHLBERG, E., AND J.-F. MERTENS: "On the Strategic Stability of Equilibrium," mimeo, Harvard Business School, 1982.
[12] KREPS, D.: "Signalling Games and Stable Equilibria," mimeo, Stanford Business School, 1984.
[13] KREPS, D., P. MILGROM, J. ROBERTS, AND R. WILSON: "Rational Cooperation in the Finitely-Repeated Prisoner's Dilemma," *Journal of Economic Theory*, 27(1982), 245–252.
[14] KREPS, D., AND R. WILSON: "Reputation and Imperfect Information," *Journal of Economic Theory*, 27(1982), 253–279.
[15] ———: "Sequential Equilibria," *Econometrica*, 50(1982), 863–894.
[16] LOCKWOOD, B.: "Perfect Equilibria in Repeated Games with Discounting," mimeo, Birkbeck College, 1983.
[17] MILGROM, P., AND J. ROBERTS: "Limit Pricing and Entry Under Incomplete Information," *Econometrica*, 50(1982), 443–460.
[18] RADNER, R., R. MYERSON, AND E. MASKIN: "An Example of a Repeated Partnership Game with Discounting and with Uniformly Inefficient Equilibria," *Review of Economic Studies*, 53(1986), 59–70.
[19] RUBINSTEIN, A.: "Equilibrium in Supergames," Center for Mathematical Economics and Game Theory, Hebrew University of Jerusalem, 1977.
[20] ———: "Equilibrium in Supergames with the Overtaking Criterion," *Journal of Economic Theory*, 21(1979), 1–9.

□ □ □

Credit and Efficiency in Centralized and Decentralized Economies

M. DEWATRIPONT
DULBEA and ECARE (Université Libre de Bruxelles), CEPR and CORE

and

E. MASKIN
Harvard University

We study a credit model where, because of adverse selection, unprofitable projects may nevertheless be financed. Indeed they may continue to be financed even when shown to be low-quality if sunk costs have already been incurred. We show that credit decentralization offers a way for creditors to commit *not* to refinance such projects, thereby discouraging entrepreneurs from undertaking them initially. Thus, decentralization provides financial discipline. Nevertheless, we argue that it puts too high a premium on short-term returns.

The model seems pertinent to two issues: "soft budget constraint" problems in centralized economies, and differences between "Anglo-Saxon" and "German-Japanese" financing practices.

1. INTRODUCTION

We investigate how the degree to which credit markets are centralized affects efficiency when there is asymmetric information. Specifically, we argue that decentralization of credit may promote efficient project selection when creditors are not fully informed ex ante about project quality.

Our starting point is the idea that, although an entrepreneur (project manager) may have a relatively good idea of her project's quality from the outset, creditors acquire this information only later on, by which time the criteria for profitability may have changed. Thus, a poor project (one whose completion time is too long to be profitable ex ante) may nevertheless be financed, since a creditor cannot distinguish it at the time from a good (quick) project. Moreover, the project may not be terminated even after the creditor has discovered its quality, if significant sunk costs have already been incurred. If the threat of termination deterred entrepreneurs from undertaking poor projects in the first place, creditors would wish to commit ex ante not to refinance them. But, sunk costs may well render this threat incredible: ex post, both creditor and entrepreneur could be better off carrying on with the project, i.e. refinancing it.

How can decentralization help in such circumstances? We conceive of a decentralized credit market as one in which ownership of capital is diffuse, so that the capital needed to refinance a poor project may be available but not in the hands of the initial creditor. This creditor, we assume, can monitor the project and thereby enhance its value. However, monitoring is not observable to subsequent creditors. Consequently, the initial creditor's

incentive to monitor is blunted (relative to a centralized market where he owned all the capital) because he cannot fully appropriate the marginal return from doing so. With incentives reduced, he will monitor less than under centralization, which in turn reduces the value of the project and therefore the profitability of refinancing. That is, refinancing is less likely than in a centralized market; the threat to terminate a project is more credible.[1] Entrepreneurs are thereby induced not to undertake poor projects in the first place, and this enhances efficiency.[2]

Decentralization tends to deter projects that drag on too long, but for similar reasons may also discourage profitable projects that are slow to pay off. That is, the same features that strengthen commitments to terminate poor projects foster an over-emphasis on short-term profit opportunities.

To see this, suppose that the slow-and-quick-project model we have sketched is enriched so that not only poor projects but also highly profitable projects require long-term financing. Poor (i.e. inept) entrepreneurs are stuck with poor projects, but good (i.e. capable) entrepreneurs have a choice about whether their project is to be long-term and highly profitable or short-term and only moderately profitable. Finally, suppose that the degree of decentralization in the credit market is determined *endogenously*. That is, owners of capital can come together and *choose* whether to form a few big "banks" or a lot of small banks.

In such a model *multiple equilibria* (and, hence *coordination problems*) may well arise. If, in equilibrium, banks are small, even good entrepreneurs will have trouble getting continued financing for long-term projects for the reasons mentioned above. Thus they will choose the short-term option. But given that they do so, it will pay banks to be small (a single big bank would be overrun with unprofitable long-term projects from poor entrepreneurs). Thus, an equilibrium with only short-term projects and small banks exists.

But another equilibrium is also possible, one in which all banks are big. With a profusion of big banks, good entrepreneurs can get long-term financing and so choose highly lucrative projects. The profits from these projects outweigh the losses that banks incur from poor projects (which because of adverse selection are also financed). Such a "long-term" equilibrium can, in fact, be shown to Pareto-dominate the "short-term" equilibrium.

We believe that our framework may be relevant for two widely-discussed issues: the "soft budget constraint" problem of centrally-planned economies and the contrast in financing practices and investment horizons between economies of the "Anglo-Saxon" and "Japanese-German" modes.

Kornai (1979, 1980) has emphasized that the absence of bankruptcy threats in socialist economies resulted in the proliferation of inefficient enterprises. Firms realized that their losses would be covered by the state, and so operated quite independently of profit considerations. The pervasiveness of these soft budget constraints under socialism is widely acknowledged, and attempts to harden them are central features of several recent proposals for reform in eastern Europe.

But although the consequences of soft budget constraints have been intensively investigated, the same is not true of their causes. Most explanations have focused on political

1. Lack of commitment in centralized settings has been the focus of the *ratchet effect* literature. (See, for example, Freixas *et al.* (1985), Laffont and Tirole (1988), and Schaffer (1989)). What remains unsettled in this particular literature, however, is why lack of commitment should pertain particularly to centralization. Our paper attempts an answer to that question.

2. As in Stiglitz and Weiss (1981), creditors face an adverse selection problem. In the Stiglitz-Weiss model, credit rationing is a way to deal with this problem and improve the mix of projects being financed. In our setting, by contrast, it is the threat of termination that serves as the device for screening out poor projects.

constraints, such as the need to avoid unemployment or socially costly relocation. While not denying the importance of such constraints, we wish to suggest that *economic* factors may also be relevant. Specifically, the slow-and-quick-project model outlined above (and presented in detail in Sections 2 and 3) offers an explanation of soft budget constraints in which "softness" arises from the profitability of refinancing poor projects. Indeed, in our framework, softness is the "normal" state of affairs; the pertinent question is how, in some circumstances (e.g. a decentralized credit market), budget constraints can be hardened.[3]

We can also apply the framework to explain differences between Anglo-Saxon (U.S. and U.K.) and German-Japanese corporate finance. Several economists have noted that large German or Japanese firms have been more likely to obtain financing from banks than their American or British counterparts (which have relied more on equity or bonds for external finance). Moreover, these banking relationships have typically had a long-term structure in which banks assumed an active monitoring role. (See Aoki (1990), Baliga and Polak (1994), Corbett (1987), Edwards and Fischer (1994),[4] Mayer and Alexander (1989) and Hoshi, Kashyap and Scharfstein (1988, 1989).) Most important from our standpoint, the Anglo-Saxon/German-Japanese financial contrast seems to be marked by differences in project length. Specifically, German and Japanese corporations have seemed less prone to "short-termism" (see for example Corbett (1987) and *The Economist* (1990)).

Although highly stylized, the enriched model sketched above, is consistent with these differences. The "long-term" equilibrium accords with German-Japanese experience, and the "short-term" equilibrium with that of the U.S. and U.K.

We proceed as follows. In Section 2, we present a very simple (in some respects, over-simplified) model and show how credit decentralization can improve efficiency. We then discuss several alternative specifications that lead to the same conclusions. In particular, we argue that the contrast between centralization and decentralization is only heightened if we suppose, following one tradition, that the central financing authority maximizes social surplus rather than profit.

Section 2 distinguishes decentralization from centralization rather crudely by identifying the former with two creditors and the latter with one. In Sections 3 and 4 we turn to a richer model in which the market structure is determined endogenously. Section 3 establishes that the main qualitative conclusions of Section 2 carry over to a framework in which market structure is determined endogenously. Finally, Section 4 introduces profitable long-run projects as an additional option for good entrepreneurs and shows that there can be two (Pareto-ranked) equilibria marked by different average project lengths.

2. DECENTRALIZATION AS A COMMITMENT DEVICE

a. The Model

There are three periods, one entrepreneur, and either one or two creditors (banks). Contracting between the entrepreneur and a bank occurs in period 0, and projects are carried out in periods 1 and 2. If a project remains incomplete at the end of period 1, the entrepreneur and bank can renegotiate the terms of the contract to their mutual advantage.

3. Qian and Xu (1991) and Qian (1994) have used this approach to show how soft budget constraints both interfere with innovation and can contribute to the endemic shortages that plague socialism.

4. Edwards and Fischer (1994) note, however, that the reliance on external finance among German banks has not been so great as commonly supposed.

The entrepreneur's project can be either good (g) or poor (p). A good project is completed after one period; a poor project requires two periods for completion. (We identify the quality of a project with that of its entrepreneur; thus, we shall refer to good and poor entrepreneurs). The project generates an observable (and verifiable) monetary return only at its completion. Whether good or poor, it requires one unit of capital per period (all returns, capital inputs, and payoffs are denominated in money).

The entrepreneur has no capital herself and so has to obtain financing from the bank(s). Banks have capital but cannot initially distinguish between good and poor projects. Let α be the prior probability that the project is good. All parties are risk neutral, i.e. they maximize expected profit.

For the time being we will assign no bargaining power to the entrepreneur (we will relax this assumption in Section 3). Thus, in negotiating financial terms, a bank can make a take-it-or-leave-it offer to the entrepreneur and thereby extract the entire observable return. The entrepreneur is limited to unobservable private benefits such as the perquisites she can command, the enhancement of her human capital and reputation, or what she can divert from the project into her own pocket.

Let E_g be a good entrepreneur's private benefit. E_t is a poor entrepreneur's benefit when her project is terminated after the first period, whereas E_p is her benefit from a completed project. We assume that $E_p \geq E_t$. This inequality makes sense if we imagine that the entrepreneur can extract more from a project the longer it continues. It would also follow from a more elaborate model in which her reputation is enhanced if the project is completed. In any case, it must hold in any model in which poor projects are ever refinanced (provided that the entrepreneur always has the option of quitting after the first period).[5] We allow for the possibility that any of E_g, E_t, and E_p may be negative,[6] which could occur, for example, if private benefits include the cost of effort that the entrepreneur must incur to set the project up.

Consider centralization first. In this case, there is a single bank B endowed with two units of capital. In period 0, the entrepreneur E (whose type is private information) turns up and requests financing (i.e. a loan of one unit of capital). B makes a take-it-or-leave-it contract offer in which the repayment terms depend on the observable return and when it is realized (because E has no endowment, the repayment cannot exceed the observable return[7]). Assume that a good project generates observable return $R_g > 1$, which, given its bargaining power, B can fully extract (provided that $E_g \geq 0$; if $E_g < 0$, B can extract only $R_g + E_g$ because E will require an inducement $-E_g$ to undertake the project).

If the project is poor, B obtains nothing unless he agrees to refinancing at the beginning of period 2, i.e. agrees to loan another unit of capital[8] (since the observable return is zero at the end of the first period). Moreover, we assume that regardless of the first period agreement, B cannot commit himself not to refinance (or, rather, that any such commitment can be renegotiated). If refinanced, the poor project's observable return at the end of the second period is a random variable $\tilde{R}_p$, whose realization is either 0 or $\bar{R}_p$, where $0 < \bar{R}_p$. (We could allow R_g to be a random variable as well, but this would not matter in view of the parties' risk neutrality.) One can interpret $\bar{R}_p$ as the liquidation or resale value of the completed project. We suppose that, in addition to its role as lender, B serves to

5. And it is precisely the problem created by refinancing poor projects that is of interest to us.

6. As we shall see, in fact, the major case of interest for our purposes is where $E_t < 0$ and $E_p > 0$.

7. This is not necessarily true if the private return is known to be positive and bounded away from zero. But as long as B is uncertain about the value of this private return, he will not be able to extract it fully.

8. Here we are assuming for convenience that E cannot contribute any of what she may have saved from the first loan to reduce the size of the second.

TABLE 1

Payoffs under centralization

	Good project (assuming $E_g > 0$)	Poor project without refinancing	Poor project with refinancing
Entrepreneur	E_g	E_l	E_p
Bank	$R_g - 1$	-1	$\Pi_p^* - 2$

monitor the project.[9] This is modeled by assuming that, through his efforts, B can influence the distribution of $\tilde{R}_p$.[10] Assume that B learns E's type at the beginning of period 1. If E is poor, B can expend *monitoring effort* $a \in [0, 1]$ to raise the expectation of $\tilde{R}_p$. Specifically, let a be the probability of $\bar{R}_p$. As a rises, so does the cost of B's efforts. Let $\psi(a)$ denote this cost, with $\psi' > 0$, $\psi'' > 0$, $\psi(0) = \psi'(0) = 0$, and $\psi'(1) = \infty$. These assumptions ensure an optimal effort level $a^* \in (0, 1)$ such that $\bar{R}_p = \psi'(a^*)$ and, given its bargaining power, an expected return for B (gross of its capital investment) of $\Pi_p^* \equiv a^* \bar{R}_p - \psi(a^*)$.

To summarize, the payoffs (net of the cost of capital) of the entrepreneur and bank under centralization are displayed in Table 1.

Under decentralization, the model is much the same as above, but now assume that there are two banks, B_1 and B_2, each with only one unit of capital. The entrepreneur presents herself to B_1 in at the beginning of period 1 (we will postpone the issue of competition between banks until Section 3). If she turns out to be good, the analysis is as above. The same is true if she is poor but not refinanced. If, however, she is to be refinanced, she must turn to B_2, since by then B_1 has no capital left.[11] Suppose that any monitoring that B_1 has done in period 1 is unobservable to B_2.

For the sake of comparability, we assign B_2 no bargaining power so that, as in the case of centralization, B_1 can make take-it-or-leave-it offers. The problem for B_1 is to convince B_2 to loan a second unit of capital in exchange of a share of $\tilde{R}_p$. The higher B_2's expectation of B_1's monitoring effort in period 1, the smaller this share can be. We claim that equilibrium monitoring effort is less than a^* (the effort level under centralization), despite the fact that endowing B_1 with all the bargaining power maximizes his incentive to monitor. To see this, let $\hat{a}$ be B_2's assessment of the expected level of B_1's monitoring activity. Then, to induce B_2 to participate, the repayment he receives must be $1/\hat{a}$ if $\tilde{R}_p = \bar{R}_p$. This means that B_1 chooses a to maximize

$$a(\bar{R}_p - 1/\hat{a}) - \psi(a),$$

i.e. to satisfy $\bar{R}_p - 1/\hat{a} = \psi'(a)$.[12] Now, in equilibrium, $\hat{a}$ must be correct, so that if a^{**} is the equilibrium effort level, a^{**} satisfies $\bar{R}_p = \psi'(a^{**}) + 1/a^{**}$.[13] Clearly, a^{**} is less than a^* (because B_1 concedes part of the marginal return from monitoring to B_2). Therefore, $\Pi_p^{**} \equiv a^{**}\bar{R}_p - \psi(a^{**})$ is less than Π_p^*.

9. In the 1990 version of this paper, we assumed that, instead of monitoring, B acquires information about the project that it can use to affect the distribution of $\tilde{R}_p$.

10. We could also assume that monitoring affects the realization of R_g. Because such monitoring would play no role in our analysis, however, we do not consider it.

11. Actually, all that is needed for our purposes is that B_1 should not be willing or able to undertake all the refinancing itself. Indeed, even if B_1 had more than 1 unit of capital left, B_2 would still have to be brought in if B_1 were sufficiently risk averse.

12. This first-order condition is valid provided that $\bar{R}_p \geq 1/\hat{a}$. Otherwise, the maximizing choice of a is $a = 0$.

13. If there is no solution to this equation, then $a^{**} = 0$ (see footnote 12). If there are several solutions, choose the one that maximizes $a(\bar{R}_p - 1/a) - \psi(a)$, in order to rule out inefficiencies due simply to coordination failure.

TABLE 2

Payoffs under decentralization

	Good project (if $E_g > 0$)	Poor project with no refinancing	Poor project with refinancing
E	E_g	E_l	E_p
B_1	$R_g - 1$	-1	$\Pi_p^{**} - 2$
B_2	0	0	0

Recapping we exhibit the (net) equilibrium payoffs under decentralization in Table 2.

We are interested in comparing the (perfect Bayesian) equilibria under centralization and decentralization, and, especially, in investigating how these two alternatives fare in deterring poor entrepreneurs. For these purposes, it makes sense to suppose that poor projects generate negative "social surplus" ($\Pi_p^* + E_p < 2$),[14] that good projects have positive surplus ($R_g + E_g > 1$), and that poor entrepreneurs are deterred only by termination[15] ($E_l < 0 < E_p$). We shall (briefly) consider the other cases after Proposition 1 and in Section 3. (Not surprisingly, centralization and decentralization perform very similarly in most of those other cases.)

Proposition 1. *Assume that $E_p > 0 > E_l$. Under either centralization or decentralization, there exists a unique equilibrium. For parameter values such that some financing is undertaken in equilibrium, a necessary and sufficient condition for project selection to differ in the two equilibria is $\Pi_p^* > 1 > \Pi_p^{**}$. If this condition holds, only a good project is financed under decentralization (the socially efficient outcome); both good and poor projects are financed (and the latter refinanced) under centralization.*[16]

Sketch of Proof. If $\Pi_p^* < 1$, then it is inefficient to refinance a poor project under centralization (and *a fortiori* under decentralization). Thus, a poor entrepreneur will not seek financing (since $E_l < 0$), and so only a good project is financed under both centralization and decentralization. If $\Pi_p^{**} > 1$, then once even a poor project is started, parties will end up refinancing it under decentralization (and *a fortiori* under centralization). Because $E_p > 0$, we conclude that a poor entrepreneur will gain by getting funded and so, under

14. Even if $\Pi_p^* + E_p > 2$, a poor project may not necessarily be desirable. In view of the unobservability of the entrepreneur's private return, she cannot be made to compensate the centralized creditor for its negative profit $\Pi_p^* - 2$. Thus the project's desirability will depend on the creditor's and entrepreneur's relative weights in the social welfare function. However, if $\Pi_p^* + E_p < 2$, then a slow project is unambiguously inefficient.

15. Poor entrepreneurs might be threatened by legal sanctions (e.g. the threat of being thrown in jail), which could have a deterrent effect. However, if these entrepreneurs are needed for the completion of the project in the second period, such threats may not be very credible.

16. As modelled, negotiation between the entrepreneur and the bank can occur only after period 1 has elapsed, i.e., after one unit of capital has already been sunk. Let us consider what would happen if regeneration were also permitted *before* the capital is sunk (but after the initial financing contract has been signed). In that case, the bank could propose returning the first period's capital unused in exchange for a fee of $E_p + \varepsilon$. A poor entrepreneur would accept this deal, whereas a good entrepreneur would not (provided that $E_p + \varepsilon < E_g$). Moreover, given our assumption that $\Pi_p^* + E_p - 2 < 0$, the bank would be better off. To rule out such a peculiar outcome, we can suppose that, in addition to good and poor projects, there is a third type that is so dreadful that refinancing is never desirable but for which the entrepreneur's payoff is positive if financed for even one period. Let us suppose that, with high probability, the quality of such a project is detected by the bank before the capital is sunk. Nevertheless if the probability is less one, dreadful entrepreneurs will still seek financing. Therefore, the bank will thwart its detection mechanism and seriously interfere with efficiency if it offers the above deal.

both decentralization and centralization, both types of projects will be financed.[17] Finally, if $\Pi_p^{**} < 1 < \Pi_p^*$, refinancing is efficient under centralization but not under decentralization. Hence, a poor project will be funded in the former case but not the latter. ‖

Hence, either centralization and decentralization lead to the same project selection in equilibrium,[18] or else decentralization is strictly better, i.e. it selects efficiently whereas centralization is subject to a soft budget constraint.

We have been assuming that $E_t < 0 < E_p$. If $E_t > 0$, then termination does not deter a poor entrepreneur from seeking financing, and both poor and good projects are financed under either centralization or decentralization (although that is not to say that the two systems are equally efficient; see footnote 17). If $E_p < 0$, then only good projects are financed under either system.

b. Alternative Specifications

We have modeled the initial project selection as a problem of adverse selection and refinancing as one of moral hazard, but these imperfections can readily be switched around. Specifically, suppose that instead of project length being given exogenously, E can affect it through (unobservable) effort. Under centralization, B could reward the entrepreneur for early completion, but such a reward might make financing unattractive from B's perspective. The advantage of decentralization would be to induce E to complete early without having to reward her; the threat of termination would be inducement enough. Such an alternative model should yield qualitatively very similar results. Undoubtedly, both specifications are relevant in reality.

By the same token, B_2's informational disadvantage has been formally expressed as a problem of moral hazard but could alternatively be derived from adverse selection and *collusion* between E and B_1. Let us, for example, drop B_1's effort from the model (so that $\tilde{R}_p$'s distribution becomes exogenous) but also abandon the assumption that $\tilde{R}_p$'s realization is verifiable. Interpret B_1's informational advantage as the ability to prove to a court that $\tilde{R}_p = \bar{R}_p$, if that equality holds. As in models of hierarchies (Tirole (1986), Kofman and Lawarrée (1993)), collusion between two parties who share some information may prevent a third party without access to that information from sharing the benefits. Here it would be in B_1's and E's joint interest to agree to conceal the evidence that $\tilde{R}_p = \bar{R}_p$ (putting aside the unresolved theoretical issue of how such an agreement would be enforced) in order to prevent B_2 from extracting some of the return. Hence decentralization, by giving rise to collusion, reduces the incentive to refinance poor projects, as in subsection *a*.

In our model, it is the non-transferability of information that makes multi-creditor financial arrangements problematic. But there is a related (yet informal)[19] idea from the finance literature that would serve our purposes just as well: the principle that renegotiation

17. This relies on our assumption that some financing is undertaken in equilibrium. If this assumption is violated, then it is possible that no projects are financed under either system, or even that both are financed under centralization and neither under decentralization. The latter possibility is an artefact, however, of the crude way we have modelled decentralization. If the market structure is determined endogenously (as in the model of Section 3), this particular discrepancy between centralization and decentralization disappears.

18. But not necessarily the same degree of efficiency. If $\Pi_p^{**} > 1$, both centralization and decentralization select the same projects, but the former is more efficient, since $\Pi_p^* > \Pi_p^{**}$. However, this discrepancy derives from our over-simplified model of decentralization (see footnote 16). In the more satisfactory model of Section 3, centralization and decentralization are equally efficient in the case where they make the same project selection.

19. See Bolton and Scharfstein (1994) and Hart and Moore (1995) for two recent contributions that build upon this insight.

becomes more difficult to coordinate the more parties are involved. From this standpoint, having two creditors reduces the chances of refinancing because getting them to agree to it is harder.

We have endowed the creditors in both the centralized and decentralized models with the same objective: expected profit maximization. But ever since Lange and Lerner it has been common practice to have the centre in planned economy models maximize expected *social surplus*. To do so here would, in fact, only aggravate the inefficiency of centralization. To see this, recall that centralization's shortcoming is that it promotes "too much" refinancing. Now if, at the beginning of period 2, the creditor takes into account total social surplus rather than just its own profit (see footnote 14, however, for why social surplus is not unambiguously the best measure of efficiency in this model), the criterion for refinancing would become $E_p + \Pi_p^* > 1$, i.e. it would be more relaxed than before and so refinancing would occur even more readily.

3. EQUILIBRIUM IN A DECENTRALIZED CREDIT MARKET

The contracting model of the previous section is rather "microeconomic" in nature, involving a single entrepreneur and at most two creditors. For the case of centralization, assuming only a single creditor seems quite reasonable; in many centralized economies, the state has been the only significant lender. However, to equate decentralization with the existence of two banks is fairly heroic (or foolhardy). Moreover, our model leaves out two ingredients that are important features of decentralized credit markets, namely, competition among creditors and the endogenous determination of the market structure.

Thus in this section, we enrich the previous model of decentralization by assuming that there is an indefinitely large population of (identical) investors, each endowed with a small amount of capital. Thus, as in the introduction, a decentralized market is one with *diffuse ownership*, in the sense that there are many small investors. Investors, however, are allowed to *join forces* at the beginning of period 1, to form banks. Each bank has capital equal to the sum of its investors' endowments and should be viewed as a *cooperative*, i.e. as managed jointly with all members having access to the information acquired when monitoring a project. (Actually, given our risk-neutrality assumption, we could alternatively assume that joining forces entails setting up a lottery that gives each participant a chance to receive *all* the capital). But the transfer of information across banks is assumed to be impossible.

We assume that there is a population n of entrepreneurs, each drawn independently from a distribution in which there is a probability α of being good. Although n should be thought of as large, the indefinite supply of capital (which we may suppose takes the form of a liquid asset with interest rate normalized to zero) ensures that every project can in principle be financed. Operationally, this will have the effect of driving creditors' profits to zero through competition (i.e. the entrepreneur will now retain some of the observable return herself).

As for centralization, we modify the model of subsection 2a only by adopting the above assumption of n entrepreneurs and by supposing that the single creditor has enough capital to accommodate them all. The earlier analysis of equilibrium in the centralized case clearly carries over completely.

As modelled, centralization differs from decentralization in two respects: ownership of capital and transferability of information. It is this combination of attributes that generates our results. Of course, we are idealizing the quality of the flow of information within a centralized hierarchy, and our perspective is quite "un-Hayekian" in that respect.

Still this flow, however imperfect, is likely to be better than the transferability of information between separate (and competing) hierarchies (e.g. rival banks).

The timing of our modified decentralization model is as follows. At the beginning of period 1, investors can join forces to form banks (in equilibrium, not all investors need do so). An investor can contribute his capital to a bank of any "size" he chooses (because all creditors are identical and there are indefinitely many of them, he will be able to find sufficiently many other like-minded investors in equilibrium to actually form the bank). At the same time, each bank/creditor offers a set of contracts (a contract is the same as in Section 2). Entrepreneurs then choose among contracts. If more than one entrepreneur chooses the same contract, then there has to be rationing (see below). If after period 1 some projects are not yet complete, existing or new creditors can offer refinancing contracts. The affected entrepreneurs then choose among these contracts (again, possibly with some rationing).

Because everyone is risk neutral, there is no advantage to diversification *per se*, and so in equilibrium the largest creditor that need form is one with two units of capital. We shall refer to creditors with one and two units of capital[20] as small and large creditors, respectively.

Notice that what we are referring to as a bank's "size" is more accurately thought of as the bank's *liquidity*—how much of its assets are available to be loaned out—which may bear little relation to its literal size, i.e. *total* assets. Thus, the terms "small" and "large" creditor might more properly be relabelled "illiquid" and "liquid" creditor. (From this perspective, soft budget constraints arise in a centralized economy because the centre is too liquid, e.g. it can print money to refinance projects.)

A small creditor must invest all its capital in a single project if it is to do any financing in period 1. In this case the refinancing problem is the same as in subsection 2a.

A big creditor has two choices: it can fund a single project and keep its second unit of capital liquid, or it can finance two projects, thus sinking all its capital. Such a creditor is, respectively, denoted *diversified* or *undiversified* (the usage here is not quite standard because, as noted, ordinary diversification plays no role). A poor entrepreneur financed by a diversified creditor knows that its chance of being refinanced is the same as in the centralized model of subsection 2a. When the creditor is undiversified, however, refinancing possibilities depend on its *mix* of projects. Indeed, a poor entrepreneur financed by such a creditor is in the same situation, if the creditor's other project is also slow, as though financed by a small creditor. In this sense, lack of diversification is a *substitute* for being small. However, it is not a perfect substitute because the poor entrepreneur *can* obtain refinancing if the other project is good. (The creditor can use the return on the good project either to refinance the poor one directly or—if this return is realized too late—as collateral against a loan from another bank.)

As we have mentioned, entrepreneurs have to be rationed if more than one chooses the same financing contract. By a *rationing scheme* we mean a rule that, for any set of contracts that could be offered, specifies, for each contract in the set and each entrepreneur, the probability that the contract is assigned this entrepreneur. For our purposes, many different schemes would do. For concreteness, we concentrate on the following simple scheme:

20. Notice that it is of no value and possibly actually harmful to have strictly between one and two units of capital. To prevent refinancing from occurring, it is better to have one unit. And if refinancing *does* occur, it is better to have two.

The Rationing Scheme:[21] All entrepreneurs of a given type are first allocated uniformly over the set of their favourite contracts (this reflects the attempt by an entrepreneur to choose the best contract for herself); if there are fewer entrepreneurs of a given type than favourite contracts, the entrepreneurs are allocated at random to these contracts. If only one entrepreneur is allocated to a given contract, she is *assigned* to that contract with probability one. If more than one is allocated, each has an equal chance of being assigned. At the end of this round, the procedure is repeated with all entrepreneurs and contracts not yet assigned. The process continues iteratively until either the supply of unassigned entrepreneurs or that of desirable contracts (those that are preferred to no contract at all) is exhausted.

Instead of modelling entrepreneurs' behaviour explicitly, we shall subsume it within the rationing scheme (which is applied both after the period 1 and period 2 contracts are offered). We can thus define equilibrium in terms of creditors' behaviour alone.[22]

Equilibrium. An equilibrium is a configuration of creditors, each creditor's set of period 1 contracts (possibly empty), and each creditor's refinancing strategy (the period 2 contracts it offers as a function of what happened in the first period) such that, given the rationing scheme,

(i) each creditor earns non-negative expected profit on each of its contracts (whether first or second period) given other creditors' contracts and their refinancing strategies;

(ii) there is no other set of contracts that a creditor could offer and no other refinancing strategy that, given others' behaviour, would earn higher expected profit;

(iii) there is no group of inactive investors (i.e. investors who do not already form a bank) who could come together to become a creditor with a set of contracts and a refinancing strategy that, given the behaviour of the already existing creditors, makes strictly positive expected profit.

We will focus on pure-strategy equilibria (where, moreover, all creditors of a given size offer the same contracts).

As in Section 2, we are interested in comparing equilibria under centralization and decentralization. Once again, the interesting case (i.e. the case where there is a significant difference) is $E_p > 0 > E_r$, and so we shall stick to this assumption. We shall also continue to assume that $E_g + R_g > 1$ (good projects are efficient), and, that $E_p + \Pi_p^* < 2$, i.e. poor projects are inefficient (but see the discussion of equilibrium efficiency after Proposition 3, where this is relaxed).

When $\Pi_p^* > 1 > \Pi_p^{**}$, we have seen that the centralized outcome entails inefficient project selection: both good and poor projects are financed. According to the simple model of Section 2, decentralization hardens the budget constraint and induces an efficient outcome in which only good projects are funded. We now observe that the same conclusion

21. We ignore the issue of strategic behaviour on the part of entrepreneurs, i.e. the possibility that an entrepreneur will choose a less favoured contract because she has a better chance of being assigned it. However, with enough uncertainty about who the other entrepreneurs are, etc., such behaviour would not be optimal in any case.

22. Actually, it is investors, rather than creditors, who are the basic decision-making units. We find it too cumbersome, however, to define equilibrium in terms of investor behaviour. Whichever way one does it, the "natural" notion of equilibrium is not entirely clear. This is because if an investor contemplates joining a bank of a given size he must compare the corresponding payoff with what he would get if he joined some other bank. But what is he to suppose happens to the first bank if he does not join it? (The answer may well be relevant to his payoff.) That it finds a replacement for him? That it does not form at all? Implicitly, our definition of equilibrium adopts the former hypothesis.

obtains for our more elaborate model (Proposition 2 shows that an equilibrium with this hardening feature exists, and Proposition 3 demonstrates that it is essentially unique). Basically, this is because creditors would like to avoid financing poor projects. Hence, whether or not there is competition among them, they will extract all the observable surplus from such projects. And so, even in this more elaborate model, the condition $\Pi_p^* > 1 > \Pi_p^{**}$ continues to imply that refinancing will occur with big creditors but not small.

Proposition 2. *Suppose $\Pi_p^* > 1 > \Pi_p^{**}$. There exists an equilibrium in which each of $n+1$ or more small (one-unit) creditors offers a first-period contract that just breaks even on good entrepreneurs. No big creditors (two or more units) offer first-period contracts.*

Proof. Each of these small creditors earns zero profit because it breaks even on good entrepreneurs and does not attract poor entrepreneurs (since they cannot be refinanced). Moreover, none of these creditors could make positive profit by deviating because any contract that earned positive profit on good entrepreneurs would not (in view of the rationing scheme) be allocated any of them since there are enough other small creditors (that is, at least n) offering more favourable terms to accommodate all good entrepreneurs. Finally, no new creditor can enter and make positive profit: it cannot make money on good entrepreneurs for the reason just given, and if it attracted poor entrepreneurs (which would require that it consist of two or more units since $\Pi_p^* > 1 > \Pi_p^{**}$), it would lose money on them since $\Pi_p^* < 2$. ‖

Proposition 3. *If $\Pi_p^* > 1 > \Pi_p^{**}$, then the only equilibrium is that described in Proposition 2.*[23]

Proof. We first show that there cannot be an equilibrium in which a big creditor offers any first-period contracts. If there were such contracts in equilibrium, then there would be one to which a poor entrepreneur is assigned with positive probability. (A poor entrepreneur can earn a positive return only from big creditors, contracts because, since $\Pi_p^* > 1 > \Pi_p^{**}$, only these are refinanced. Indeed, if such a contract is refinanced, the entrepreneur's return is certainly positive. This will be the case when the big creditor is diversified, but also when undiversified provided that the other project financed is good. Thus, it cannot be the case that every big creditor contract is assigned only good entrepreneurs.) Of the contracts that are assigned poor entrepreneurs with positive probability, let c^0 be the one that gives poor entrepreneurs the best terms. Contract c^0 earns a negative return on poor entrepreneurs (since $\Pi_p^* < 2$), and so, in order to earn a non-negative return over all, it must earn a strictly positive return on good entrepreneurs and be assigned them with positive probability. Suppose that a group of investors who are inactive in equilibrium come together as a small creditor and offer a contract c^{00} with slightly more favourable terms for good entrepreneurs than c^0 (i.e. the contract c^{00} slightly "undercuts" c^0). This contract c^{00} must be assigned good entrepreneurs. But because it is not refinanced (since $\Pi_p^{**} < 1$) it will not be assigned poor entrepreneurs. Therefore, it makes positive profit overall, a contradiction. We conclude that big creditors cannot offer first-period contracts in equilibrium.

23. Actually, Proposition 2 describes a *multiplicity* of equilibria in which the number of active creditors can vary as long as it exceeds $n+1$. However, this sort of non-uniqueness is clearly not essential.

We next observe that the only contract that is accepted with positive probability in equilibrium is the break-even contract for good entrepreneurs. A contract that offered more favourable terms to good entrepreneurs would lose money, and a less favourable contract would make positive profit and so induce entry and slight undercutting as above.

Finally, there must be at least $n+1$ small creditors offering the break-even contract in equilibrium. Otherwise, a small creditor could enter and offer a contract that, if assigned to a good entrepreneur, would make a profit (and also would be preferred by the entrepreneur to no contract at all). Because there are fewer than n other small creditors, there would be a positive probability that not all good entrepreneurs could find financing elsewhere and therefore would be assigned this contract. ‖

The proof of Proposition 3 is somewhat involved, but the idea that underlies it is very simple: If $\Pi_p^* > 1 > \Pi_p^{**}$, then small creditors have the advantage over their big counterparts of not attracting poor entrepreneurs. Thus they are more efficient and so, in equilibrium with free entry, drive the big creditors out of the market.

We have been considering the case in which $E_p + \Pi_p^* < 2$. If instead this inequality goes the other way (but all other inequalities remain the same, in particular $\Pi_p^* < 2$), then, according to the criterion of social surplus, slow projects are efficient (see footnote 14, however, for why social surplus may not be the right criterion). Nonetheless, Propositions 2 and 3 continue to hold. That is, only good projects are financed under decentralization. This follows because creditors ignore the entrepreneurs' private benefits in deciding whether or not to fund a project, and suggests that there may be an excessive tendency in decentralized credit markets to focus on short-term (i.e. one-period) projects (because banks are too illiquid to make efficient loans). For a less ambiguous illustration of this tendency (one that does not rely on this questionable measure of efficiency), see the next section.

In the case $\Pi_p^* > 1 > \Pi_p^{**}$, the market outcome reproduces the features of the decentralized model of subsection 2a. When $\Pi_p^{**} > 1$, matters are more complicated because both types of entrepreneurs will be financed regardless of the size of creditors. A potential problem of non-existence of equilibrium may arise if a creditor is able to affect its mix of entrepreneurs (the relative probabilities of good and poor entrepreneurs choosing its contracts) sharply by slightly changing the terms it offers. Such a problem is similar to those arising in insurance models à la Rothschild–Stiglitz (1976) and Wilson (1977). To avoid all this, we introduce the following mild assumption:

Assumption A. Slow entrepreneurs have an (arbitrarily) small probability of completing their projects in one period.

This assumption limits the effect that improving the terms offered to good entrepreneurs has on a creditor's mix of entrepreneurs; any improvement will be attractive to poor as well as to good entrepreneurs. Assumption A enables us to derive the following result:

Proposition 4. *Let* $\Pi_p^{**} > 1$, *Under Assumption A, there is a unique equilibrium* (*where uniqueness is qualified the same way as in Proposition* 3) *in which* (i) *only big creditors are active in the market, and* (ii) *at least* $n+1$ *of them offer contracts that break even on average across good and poor projects and that extract the entire observable return from poor projects.*

Proof. We will show that the behaviour described constitutes an equilibrium. Uniqueness can be established as in the proof of Proposition 3. Clearly, it is not optimal

to leave any observable return to poor entrepreneurs: a creditor would only improve its mix of entrepreneurs by lowering the return offered to poor ones. Under Assumption A, however, a creditor cannot improve its mix by offering better terms to good entrepreneurs, since such improvement would attract all the poor entrepreneurs as well. Therefore, if at least n other big creditors offer break-even contracts as described in the proposition, a big creditor can do no better than to follow suit.

As for small creditors, they cannot avoid attracting poor as well as good entrepreneurs since $\Pi_p^{**} > 1$. However, they are less efficient in monitoring poor projects than are big creditors. Thus if the latter creditors break even, the former lose money. ‖

Thus, in this model, a decentralized market leads to efficient creditor liquidity. When neither large nor small creditors can commit not to refinance poor entrepreneurs, large creditors are more efficient because they have the incentive to provide better monitoring. Therefore, they drive small creditors out of the market.

4. DECENTRALIZATION AND SHORT-TERMISM

We now introduce a third project: a two-period but *very profitable* undertaking denoted by the subscript v. This project requires one unit of capital per period and generates a return $R_v > 2$ after two periods. For simplicity, we suppose that R_v is deterministic and does not require monitoring. A good entrepreneur can choose between a good or very profitable project (poor entrepreneurs are stuck with poor projects), but her choice is unobservable.[24] Moreover, poor and very profitable projects are indistinguishable to creditors at the end of period 1.[25] A good entrepreneur's private benefit from a very profitable project is E_l (if the project is terminated after one period) or E_v (if the project is completed). We adopt the natural assumption that $E_v \geq E_p$.

The timing is much the same as that of Section 3. But we now must insert the choice between good and very profitable contracts, which we assume is made at the same time as creditors offer contracts. A creditor's monitoring intensity depends on its beliefs in period 1 about project quality. Specifically, a large creditor will expend effort $a^*(\alpha')$ such that $(1-\alpha')\bar{R}_p = \psi'(a^*(\alpha'))$ if it believes that α' is the probability, the project is very profitable and $1-\alpha'$ is the probability it is poor. Note that if $\alpha' = 0$, the model reduces to that of Section 3. Hence $a^*(0) = a^*$, and the financial return (gross of capital) is Π_p^*. Similarly, for a small creditor, we define $a^{**}(\alpha')$, and obtain $a^{**}(0) = a^{**}$, which generates gross financial return Π_p^{**}. Refinancing decisions clearly also depend on α'. Pessimistic beliefs (i.e. low values of α') lead to *short-termism*—i.e. the choice of good over very profitable projects—because good entrepreneurs forecast that long-term projects will not be refinanced:

Proposition 5. *If $\Pi_p^{**} < 1$ there exists an equilibrium in which only small creditors are active and only good projects are chosen.*[26]

24. We thank Ian Jewitt for suggesting that we replace our earlier adverse selection treatment of very profitable long-run projects with the current moral hazard formulation.

25. To simplify analysis, however, we suppose that these projects *are* distinguishable at the end of period 2.

26. As the model stands, this result depends to some extent on the timing. If good entrepreneurs choose projects *before* creditors move, nothing is changed since the entrepreneurs' decisions are unobservable anyway. But if the creditors move first, then a group of investors might form a bank so big that good entrepreneurs are encouraged to choose very profitable projects. Still, the bank may have to be very big indeed—big enough to accommodate a large fraction of all entrepreneurs—otherwise, a good entrepreneur may face too high a risk of not being assigned to one of this bank's contracts if she chooses the very good project. Thus, if there are reasonable limits on creditor size/liquidity, our results should not be very sensitive to timing after all.

Proof. Suppose that $n+1$ or more small creditors are active (and no other creditors are) and offer the contract that breaks even on good projects. Suppose, furthermore, that creditors believe that, if a project has to be refinanced, then with high probability it is poor. Under these circumstances, all good entrepreneurs will choose good projects, since $\Pi_p^{**}<1$ and the creditors' pessimistic beliefs together imply that two-period projects will not be refinanced. Hence the creditors' beliefs are justified. Now, a small creditor clearly cannot do better than break even. Suppose then that a big creditor enters. It will attract *all* the poor entrepreneurs and only its share of good projects. But since the former are unprofitable ($\Pi_p^*<2$), the creditor will lose money on average. ‖

The equilibrium of Proposition 5 can be highly inefficient. As in Section 2, let α be the fraction of entrepreneurs who are good. Notice that the Proposition 5 equilibrium exists no matter how close α is to 1. However if R_b is big, then, for α near 1, it is clearly better from a social standpoint to put up with poor projects for the sake of the very good ones. Indeed, for big enough R_b, there exists another and more efficient equilibrium, provided that α is sufficiently near 1. If $E_b=E_g$, the precise condition we require for existence of this other equilibrium is

$$\alpha R_b+(1-\alpha)a^*(\alpha)\bar{R}_p-\psi(a^*(\alpha))-2>\alpha(R_g-1). \qquad (*)$$

Condition (*) implies that if all good entrepreneurs choose very profitable projects, big creditors can offer them better terms than on good projects, while still breaking even. In such a case, creditors' *optimistic* expectations are self-fulfilling.

Proposition 6. *Suppose that $\Pi_p^{**}<1$, $E_b=E_g$, and (*) is satisfied. Then there exists an equilibrium in which only big creditors form and all good entrepreneurs select very profitable projects.*

Proof. Suppose that there are at least $n+1$ big creditors and each offers the contract $\hat{c}$, which gives the entrepreneur nothing (except her private return) if the project turns out to be poor, T_b if the project is very profitable, and T_g if the project is good where

$$\alpha(R_b-T_b)+(1-\alpha)a^*(\alpha)\bar{R}_p-\psi(a^*(\alpha))-2=0 \qquad (1)$$

and

$$R_g-1-T_g=0. \qquad (2)$$

From (1), $\hat{c}$ just breaks even if creditors' beliefs that all good entrepreneurs choose very profitable projects are correct. Now, good entrepreneurs will choose these projects provided that

$$E_b+T_b>E_g+T_g. \qquad (3)$$

From (1) and (2) and because $E_b=E_g$, (3) can be rewritten as

$$\alpha R_b+(1-\alpha)\Pi_p^*-2>\alpha(R_g-1),$$

which is just (*). Hence, good entrepreneurs will select very profitable projects as claimed. The arguments that no big creditor can do better by deviating and that any creditor can profit from entering are the same as in the proof of Proposition 4. ‖

Propositions 5 and 6 imply that the same economy may end up in two quite different equilibria. In the equilibrium of Proposition 5, creditors are small and projects are short-term. In that of Proposition 6, creditors are big and projects are long-term. Note that,

even ignoring entrepreneurs' private benefits, (*) implies that the latter equilibrium is more efficient. Including the private benefits only aggravates the discrepancy (it would entail adding $(1-\alpha)E_p$ to the left-hand side of (*)). Indeed, the equilibrium of Proposition 6 Pareto-dominates that of Proposition 5.

To conclude, let us note that Propositions 5 and 6 have some connection with those of von Thadden (1995). Von Thadden argues that a commitment not to refinance projects may be an optimal screening device for creditors facing an adverse selection problem, even though it can induce short-termism on the part of good entrepreneurs. Although the set of technological opportunities available to entrepreneurs and the initial asymmetry of information in his paper are similar to those in our model, von Thadden takes a different perspective, since he does not explicitly address ex post incentives to refinance or the role of creditor liquidity. Rather, he concentrates on a one-creditor problem. In his model, bank finance can reduce short-termism thanks to economies of scale (à la Diamond (1984)), which make direct inspection of project types profitable.

Acknowledgements. We thank Patrick Bolton, Jeremy Edwards, Ian Jewitt, Charles Kahn, Janos Kornai, Colin Mayer, John McMillan, John Moore, Yingyi Qian, Gérard Roland, David Scharfstein, Chenggang Xu, and two referees for useful comments. This research was supported by the NSF, and by the Belgian Government under PAI grant No. 26.

REFERENCES

AOKI, M. (1990), "Toward an Economic Model of the Japanese Firm", *Journal of Economic Literature*, **28**, 1–27.
BALIGA, S. and POLAK, B. (1994), "Credit Markets and Efficiency" (mimeo).
CORBETT, J. (1987), "International Perspectives on Financing: Evidence from Japan", *Oxford Review of Economic Policy*, **3**, 30–55.
DIAMOND, D. (1984), "Financial Intermediation and Delegated Monitoring", *Review of Economic Studies*, **51**, 393–414.
The Economist (1990), "Punters or Proprietors?: A Survey of Capitalism", May 5–11.
EDWARDS, J. and FISCHER K. (1994) *Banks, Finance and Investment in Germany* (Cambridge: Cambridge University Press).
FREIXAS, X., GUESNERIE, R. and TIROLE, J. (1985), "Planning under Incomplete Information and the Ratchet Effect", *Review of Economic Studies*, **52**, 173–192.
HOSHI, T., KASHYAP, A. and SCHARFSTEIN, D. (1988), "Corporate Structure, Liquidity, and Investment: Evidence from Japanese Industrial Groups" (mimeo).
HOSHI, T., KASHYAP, A. and SCHARFSTEIN, D. (1989), "Bank Monitoring and Investment: Evidence from the Changing Structure of Japanese Corporate Banking Relationships" (mimeo).
KOFMAN, F. and LAWARÉE, J. (1989), "Collusion in Hierarchical Agency" (mimeo).
KORNAI, J. (1979), "Resource-Constrained versus Demand-Constrained Systems", *Econometrica*, **47**, 801–819.
KORNAI, J. (1980) *The Economics of Shortage* (New York: North-Holland).
LAFFONT, J. J. and TIROLE, J. (1988), "The Dynamics of Incentive Contracts", *Econometrica*, **56**, 1153–1175.
MAYER, C. and ALEXANDER, I. (1990), "Banks and Securities Markets: Corporate Financing in Germany and the UK" (mimeo).
QIAN, Y. (1994), "A Theory of Shortage in Socialist Economies Based on the Soft Budget Constraint", *American Economic Review*, **84**, 145–156.
QIAN, Y. and XU, C. (1991), "Innovation and Financial Constraints in Centralized and Decentralized Economies" (mimeo, London School of Economics).
ROTHSCHILD, M. and STIGLITZ, J. (1976), "Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information", *Quarterly Journal of Economics*, **90**, 629–649.
SCHAFFER, M. (1989), "The Credible-Commitment Problem in the Center-Enterprise Relationship", *Journal of Comparative Economics*, **13**, 359–382.
STIGLITZ, J. and WEISS, A. (1981), "Credit Rationing in Markets with Imperfect Information", *American Economic Review*, **71**, 393–410.
TIROLE, J. (1986), "Hierarchies and Bureaucracies: On the Role of Collusion in Organizations", *Journal of Law, Economics and Organizations*, **2**, 181–214.
VON THADDEN, E. L. (1995), "Bank Finance and Long-Term Investment", *Review of Economic Studies*, **62**, 557–575.
WILSON, C. (1977), "A Model of Insurance Markets with Incomplete Information", *Journal of Economic Theory*, **16**, 167–207.

□ □ □