Procedurally Rational Experimentation in Infinite-Horizon Games - Summary for TARK VIII

Ran Spiegler
School of Social Science
Institute for Advanced Study
Princeton NJ 08540, USA
E-mail: rani@ias.edu

May 11, 2001

1 Introduction

Standard notions of equilibrium in infinite-horizon interactions support a course of action by the players' beliefs that a deviation will be "punished". Along the equilibrium path, players optimize with respect to their beliefs and thus never realize these punishments. In the case of infinitely repeated games with complete information, this approach often results in "Folk Theorem". Under the solution concept of Nash Equilibrium (NE), for example, every individually rational payoff profile can be approximated arbitrarily closely by a pure-strategy NE, by properly fixing the punishments that sustain equilibrium behavior.

This summary proposes an equilibrium concept, which departs from NE in that along the equilibrium path, players do not strictly conform to optimizing behavior. Their departure from best-replying is not arbitrary or random, but *systematic.* It is based on the following behavioral principle: *Every threat that affects optimal response is tested* (finitely many times). Not every threat is realized, but only the "relevant" ones, whose existence affects optimal response.

To illustrate this principle, consider the following two examples:

1. A parent tries to educate his child to clean his room, using the following policy: the infant is taken to the movies on weekends if and only if he cleaned his room during the preceding week. Cleaning the room is costly for the kid but this cost is outweighed by the benefit of being taken to the movies. Best-replying implies that the child will never fail to clean his room. According to our behavioral principle, he will test his parent's policy - he will leave his room untidy finitely many times before settling down on the optimal response (cleaning his room regularly). But - if the parent did not condition any benefit on the state of the child's room, the child would never experiment with suboptimal play (i.e., with cleaning his room).

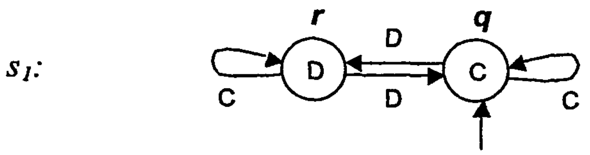
2. Leaders of two countries negotiate over an arms treaty. At a certain point in the negotiations, the leader of Country A expects that the leader of Country B will not make any concession as long as she does not concede first. When delay is costly, best-replying implies that the leader of Country A would concede immediately. Our behavioral principle implies that she will test her opponent's "tit-for-tat" threat and delay her concession for a while. But - if the Country B's decision whether to concede were independent of Country A's concessions, the latter would never experiment with suboptimal play (i.e., with making concessions).

Let us illustrate the application of this principle in the context of infinitely repeated games with discounting.[1] Consider the infinitely repeated Prisoner's Dilemma, where the players' discount factor is arbitrarily close to 1 and the stage game is given by:

	C	D
C	3,3	0,4
D	4,0	1,1

[1] I.e., player j evaluates play paths by the discounted sum of his periodic payoffs: $U_j(a^0, a^1, \ldots) = \sum_{t=0}^{\infty} \delta^t \cdot u_j(a^t)$, where $\delta \in (0,1)$ is the discount factor and $u_j(a^t)$ is player j's periodic payoff from the action profile in period t.

Figure 1 contains a visualization of a strategy of player 1 (using finite automata representation) and a play path, which is consistent with this strategy. By the payoff structure, best-replying prescribes C (D) for player 2 whenever player 1 is in the "cooperative" ("defective") state q_c (q_d). However, in the former case, optimal response is justified by player 1's threat to punish player 2 if he plays D, whereas in the latter case, D would be the optimal for player 2 even if playing C led to the same continuation as playing D. Our behavioral principle implies that player 2 will test the threat that justifies playing C when player 1 is in q_c, yet will stick to best-replying whenever player 1 is in q_d. As can be seen in Figure 1, player 2's behavior in the play path obeys to the behavioral principle.



States :	q	q	q	r	q	q	q	q ...
Player 1:	C	C	C	D	C	C	C	C ...
Player 2:	C	C	D	D	C	C	C	C ...

Figure 1

A strategy profile satisfying that along the induced path, both players' conform to the behavioral principle with respect to each other's strategy, is an "*Experimental Equilibrium*" (EXE). Thus, the basic difference between (pure-strategy) NE and EXE can be articulated by the following table:

NE	Threats are never tested in eq.
EXE	*Relevant* threats are tested (finitely many times) in eq.

Note that in the case of infinitely repeated 2×2 games, the behavioral principle can be rephrased as follows: *If optimal response requires abandon-*

ing the stage-game best reply action in a certain situation, then this action will be experimented with (finitely many times) in that situation - where a "situation" from the point of view of one player corresponds to the class of histories in which his opponent's strategy is in the same "state".

Some of the possible interpretations of EXE are hinted by the two examples given earlier. In the first example, the child may recognize the implications of long-run best-replying, but he is tempted by the myopic action. He can settle down on long-run best-replying only after having actually experienced the implications of sub-optimally myopic behavior. Thus, *self-control* considerations determine the systematic departure from best-replying. In the second example, the leader of Country A may deliberately delay her concession because of *justifiability* considerations - it is easier for her to justify her concessions, once the opponent's tit-for-tat threat has been actually realized.

Under both interpretations, the players' knowledge of each other's strategy is the same as in NE, which is based on strict optimizing behavior. Alternative interpretations that ground the player's experimentation in a genuinely imperfect state of strategic knowledge - and thus relate them to the idea of learning in games - seem to be inconsistent with the idea of equilibrium behavior. Constructing a learning-based interpretation of EXE remains a (difficult) challenge for future research.

The remainder of this summary proceeds as follows. Section 2 presents the formal definition of EXE and discusses some of its immediate implications. Section 3 characterizes a subclass of equilibria in repeated symmetric 2×2 games. It turns out that a certain domain of strategies, which generates a Folk Theorem when subjected to NE, implies a considerably more informative characterization under EXE. Section 4 concludes.

2 A Formal Definition of EXE

Consider an infinitely repeated, simultaneous-move, two-person finite game with discounting. Player j's stage-game action set is A_j. Denote player j's stage-game payoff function by u_j. Denote j's opponent by $-j$. Stage-game indifferences are assumed away. For any action $a \in A_j$ by player j, denote her opponent's stage-game best-reply by $br_{-j}(a) \in A_{-j}$. This action is also referred to as $-j$'s *myopic* action against a.

We will only consider repeated-game strategies that admit a finite automata representation $(Q_j, q_j^0, f_j, \tau_j)$, where:

1. Q_j is a finite set of states.

2. q_j^0 is the initial state.

3. $f_j : Q_j \rightarrow A$ is an output function, which specifies the action taken by player j when he is in state $q \in Q_j$.

4. $\tau_j : Q_j \times A_{-j} \rightarrow Q_j$ is a transition function, which specifies the state to which the automaton switches from state $q \in Q_j$ when the opponent plays $a \in A_{-j}$ against q.

Let $z(s_1, s_2) = ((a_1^k, a_2^k))_{k=0,1,2,...}$ be the play path induced by the strategy profile (s_1, s_2), where $a_j^k \in A_j$ is player j's action at period k. I.e., $a_j^0 = f_j(q_j^0)$, $a_j^1 = f(\tau_j(q_j^0, a_{-j}^0))$, and so forth. Given $z(s_1, s_2)$, the function $p_j : \{1, 2, 3, ...\} \rightarrow Q_j$ signifies the state of player j's strategy at period k along $z(s_1, s_2)$.[2]

It should be emphasized that in contrast to the literature following Rubinstein (1986), the finite automata formalism is not employed here in order to capture complexity considerations, but simply as a highly convenient way to articulate the idea of systematic testing of threats as part of equilibrium behavior.

Because preferences satisfy the discounting criterion, player j's strategy s_j induces a well-defined correspondence $BR_{-j} : Q_j \rightarrow A_{-j}$, which assigns to every $q \in Q_j$ the set of $-j$'s actions that are consistent with best-replying to s_j at any period k for which $p_j(k) = q$. For expositional simplicity only, assume that $BR_{-j}(q)$ is a *singleton* for every $q \in Q_j$. In words, $BR_{-j}(q)$ is player $-j$'s "best-reply" action against q, or "the right thing to do" in the "situation" q. Note that $BR_{-j}(q) = br_{-j}[f_j(q)]$ whenever $\tau_j(q, \cdot)$ is constant: if j's continuation is independent on $-j$'s move, then $-j$'s long-run and myopic best-reply actions coincide.

Given a strategy profile (s_1, s_2), for every $q \in Q_j$, we can count the number of periods k along $z(s_1, s_2)$, in which $p_j(k) = q$ and $a_{-j}^k \neq BR_{-j}(q)$ - i.e., the number of times the best-reply action was *not* played against q along $z(s_1, s_2)$. I denote this number by $e(q)$.

[2] For example, the strategy s_1 that appears in Figure 1 is represented by: $Q = \{q, r\}$; $q^0 = q$; $f(q) = C$ and $f(r) = D$; and finally, $\tau(q', C) = q'$ and $\tau(q', D) = Q/\{q'\}$ for every $q' \in Q$. Along the play path given in Figure 1, $p_1(4) = r$ and $p_1(k) = q$ for every $k \neq 4$.

Let us now formulate an equilibrium notion, in which players depart systematically from best-replying for experimentation purposes, following a simple rule of thumb: Sub-optimal play is experimented with in a certain "situation" along the play path if and only if in that situation, the long-run best-reply action do not coincide with the myopic action (where two periods k and k' belong to the same "situation" from player $-j$'s point of view if $p_j(k) = p_j(k')$). This rule is induced by the behavioral principle: A non-myopic action can be optimal only in the face of a threat to punish the myopic action and therefore, this threat needs to be tested.

Definition 1 (EXE) *(s_1, s_2) is an* Experimental Equilibrium *(EXE) if for every state $q \in Q_j$ that is visited along $z(s_1, s_2)$:*

1. $e(q) < \infty$
2. $e(q) > 0 \quad \Longleftrightarrow \quad BR_{-j}(q) \neq br_{-j}[f_j(q)]$

The first condition is that a sub-optimal action is experimented with only *finitely* many times. The second condition articulates the above experimentation rule. The way in which EXE departs from NE can be understood by considering what would be the analogue of NE in our formalism: (s_1, s_2) is a NE if $e(q) = 0$ for every $q \in Q_j$, $j \in \{1, 2\}$.

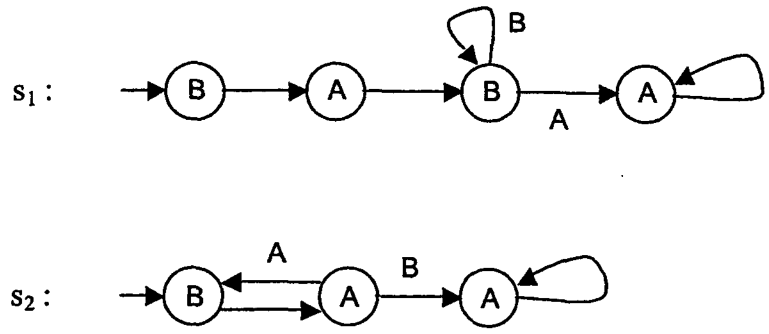
As EXE is defined in terms of pure strategies, existence is not guaranteed. However, existence of EXE is guaranteed as long as the *stage game* has a pure-strategy NE. Since EXE is a profile of strategies with a finite automata representation, it is known that the induced path eventually enters a cycle (see Osborne and Rubinstein (1994, Ch. 8)). I.e., there exist a period k^* and a cycle length L, such that $p_j(k) = p_j(k + L)$ for every $j \in \{1, 2\}$, $k \geq k^*$. Since $e(q) < \infty$, this means that all the experimentation activity necessarily takes place in the pre-cyclic phase (i.e., prior to k^*). Players conform to best-replying in the cyclic phase: $a_j(k) = BR_j[p_{-j}(k)]$ for every $k \geq k^*$.

What is the intersection between the classes of EXE and NE in repeated games? First, note that every NE that induces an (eventually cyclic) play path consisting of stage-game NE is also an EXE. Thus, pure defection in the repeated Prisoner's Dilemma, or perfect coordination in a repeated coordination game, are EXE-sustainable. The reason is that in the induced path, players' actions are consistent with both long-run and myopic best-reply actions, so there is no reason for them to experiment. This class of strategy profiles constitutes the intersection between NE and EXE. Every other EXE contains sub-optimal behavior and is therefore not a NE, while every other NE contains non-myopic behavior that is sustained by untested threats and is therefore not an EXE.

Let us construct an EXE which is not a NE. Let the stage game be a 2×2 pure coordination game:

	A	B
A	2,2	0,0
B	0,0	1,1

For sufficiently patient players, the strategy profile given by Figure 2 constitutes an EXE (the induced path is shown as well). The reader can verify that players' behavior along the play path conforms to the behavioral principle. Players 2 and 1 experiment with myopic behavior at periods 2 and 3, respectively. These experiments back-up the players' optimal, non-myopic behavior at period 4.



Player 1: B A B B A A A ...
Player 2: B A B A A A A ...

Figure 2

The play pattern has a natural interpretation. The players aim at switching from inefficient coordination to efficient coordination. For some reason, both players acknowledge that the transition is to be "led" by player 2 (and hence, it involves miscoordination). However, they first try a "shortcut" and make an instantaneous transition. Only when the shortcut results in inefficient coordination, the players go through the period of miscoordination that is necessary to arrive at efficient coordination.

I conclude this section with a few more comments on the equilibrium concept:

1. The only information on preferences that is required to define and apply EXE is the players' BR functions, defined for every possible strategy of the opponent.[3] These functions specify "the right thing to do under every situation". Once these are given, we need no more information on preferences. The only role of utility maximization in EXE is to induce well-defined BR functions.

2. Note that the definition of EXE is silent over two aspects of the players' experimentation activity:

   (a) Apart from the implication that experimentation takes place in the pre-cyclic phase of the play path, we do not know anything about the timing of experimentation. E.g., when $e(q) > 0$ for some $q \in Q_j$, does player $-j$ experiments with $-BR_{-j}(q)$ as soon as q is visited (i.e., at the earliest period k for which $p_j(k) = q$)? The definition of EXE is silent over this matter.

   (b) When the stage-game action set consists of more than two actions, the definition of EXE does not say how experiments are distributed among the sub-optimal actions. E.g., it allows players to experiment with all of them, or to experiment only with the myopic action. Obviously, this distinction is meaningless in repeated 2 × 2 games.

3. Observe that in the definition of EXE, optimality is a property of actual actions along the induced play path, not of the strategies themselves. It is therefore more appropriate to think of s_j as a *belief* held by his opponent $-j$, rather than as a conscious choice by j, following the convention of Aumann (1987) and Rubinstein (1991).

[3] The br functions can be easily embedded in BR functions.

3 EXE in Repeated Symmetric 2×2 games

This section exercises the application of EXE to repeated symmetric 2×2 games. It will be shown that a domain of strategy profiles, which is sufficiently wide to generate a Nash Folk Theorem, gives rise to a much more restrictive characterization of experimental equilibria. First, let us introduce a bit of notation. Define the min-max action $a^* = \arg\min_{a \in A_{-j}} [\max_{b \in A_j} u_j(b, a)]$. E.g., the min-max actions in the Prisoner's Dilemma and "Chicken" are "defect" and "hawk", respectively. For every action $a \in A$, denote $-a = A/\{a\}$.

In the equilibrium strategies depicted by Figure 2, the punishments that sustain the optimality of non-myopic behavior at period 4 possess a simple "trigger" structure. Playing sub-optimally myopically against a state q is punished by a number of periods (possibly zero), in which the opponent imposes the min-max outcome, before returning to q. Let us formally define this structure.

Definition 2 (Prison trigger strategies) *A strategy with finite automata representation $(Q_j, q_j^0, f_j, \tau_j)$ in a repeated symmetric 2×2 game is a* "prison" trigger strategy *if for every $q \in Q_j$ with non-constant $\tau_j(q, \cdot)$, playing $-BR_{-j}(q)$ against q is punished by $m(q) \geq 0$ periods, in which player j plays a^* before returning to q.*[4]

This is one of the simplest conceivable punishment structures in the context of repeated games. It is based on the familiar notion of "trigger" punishments - deviating from best-replying triggers a simple punishment that consists of repeating an adverse outcome for some time. In contrast to "Grim" trigger strategies, "prison" trigger strategies are "forgiving".[5] The "duration of punishment" $m(q)$ can vary with q but the player's behavior during the

[4] More formally, whenever $\tau_j(q, \cdot)$ is non-constant, Q_j contains a sequence of states $(q_n)_{n=0,\ldots,m(q)+1}$, such that:

1. $q_0 = q_{m(q)+1} = q$
2. $q_1 = \tau_j[q, -BR_{-j}(q)]$
3. $f(q_n) = a^*$ and $q_n = \tau_j(q_{n-1}, \cdot)$ for every $n \neq 0, m(q)$.

[5] Figurative speaking, the errant player ends his "imprisonment" in the exact same situation in which he originally "commited the crime" and is given another chance to behave properly.

punishment phase does not affect its duration. Thus, for instance, "Tit-for-Tat" is not a prison strategy.

Restricting attention to prison strategies is innocuous as far as NE is concerned. Every (eventually cyclic) NE-sustainable play path is also sustainable by NE with prison strategies. The idea is simple: whenever $f(q) = a^*$, we can fix $\tau(q, -BR(q)) = q$ and whenever $f(q) \neq a^*$, we can fix an arbitrarily high $m(q)$.

We will now see that the set of EXE with prison strategies is considerably more restricted.

Theorem 3 *A non-NE profile of prison trigger strategies* (s_1, s_2) *is an EXE in repeated symmetric* 2×2 *games only if:*

1. *The stage game is a coordination game with a Pareto-dominant outcome.*
2. *Along the cyclic phase of the induced play path, miscoordination occurs no more than once per cycle.*

Thus, the requirement that relevant threats are tested along the equilibrium path implies that in the case of repeated Prisoner's Dilemma, "Chicken" or "Battle of the Sexes", it is impossible to sustain non-myopic behavior with prison trigger strategies. The question of whether a more intricate structure of punishment can accomplish this is addressed at the end of this section.

To get the idea of the proof, let us focus on play paths that include periods of "mutually non-myopic" behavior - i.e., periods in which both players play non-myopically against each other. In addition, let us consider only prison strategies, in which every state is reachable from every other state - that is, every state is either visited in the cycle or can be reached from a state that is visited in the cycle.[6] Extending the proof beyond this special case is more cumbersome but does not involve genuinely new ideas.

Consider the earliest period k for which $a_j^k \neq br_j(a_{-j}^k)$ for both $j = 1, 2$. Denote $a_j^k = a_j$. By definition of EXE, for each player j there must be a period $k_j \neq k$, such that $p_j(k_j) = p_j(k)$ and $a_{-j}(k_j) = br_{-j}(a_j^k)$. But it can be shown that k must already belong to the cyclic phase of the play

[6] The strategy given by Figure 1 satisfies this condition, whereas the equilibrium strategies in Figure 2 violate it.

path.[7] Therefore, $k_1, k_2 < k$. Let $k_1 < k_2$, without loss of generality. Since k is the earliest period of mutually non-myopic behavior, the move sequence preceding k must look like this:

Period	k_1	...	k_2	...	k
Player 1	a_1	...	$-a_1$	...	a_1
Player 2	$-a_2$	...	a_2	...	a_2

By the structure of prison strategies, $-a_1 = a^*$ and player 2 is "being punished" between $k_1 + 1$ and $k - 1$ (inclusive). By EXE and the structure of prison strategies, he adheres to optimizing behavior during this punishment phase. Therefore, $a_2 = br_2(-a_1)$. Since by assumption, $a_2 = -br_2(a_1)$, the stage-game is a coordination game. It is also quite straightforward to see that $(a_1, -a_2)$ is necessarily a Pareto-dominant outcome in the stage game. Finally, since k is already part of the cycle, there can be no more than a single period of miscoordination per cycle.

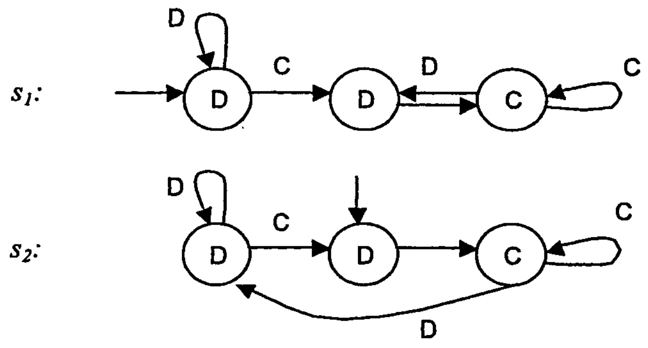
The intuition for the result is that the structure of prison strategies forces the timing of the players' experiments to be more-or-less simultaneous. At the same time that one player is punished for a previous experiment, his opponent is experimenting and getting punished for it from his own point of view. Given that players are punished by the min-max outcome, most stage-game payoff structure have to be ruled out.

The theorem also implies that in EXE of the repeated coordination game, any period of miscoordination prior to the cyclic phase (including the first period of the cyclic phase) is preceded by the same kind of play pattern we have noted in context of Figure 2. Players first try efficient coordination, which leads to a few periods of inefficient coordination, which in turn lead to miscoordination. Of course, the concept of EXE cannot explain why players find miscoordination necessary for optimal response in the first place. However, it does characterize the play patterns that accompany miscoordination.

We have seen that it is impossible to sustain cooperation in the repeated prisoner's Dilemma or Chicken with prison strategies. We have already noted that the Tit-for-Tat strategy does not belong to this class. It can be shown that in the case of the Prisoner's Dilemma, a natural generalization of the idea of reciprocity embodied in Tit-for-Tat provides a scheme of strategies,

[7] This is where the assumption that every state is reachable from every other state is used.

with which it is possible to sustain play paths, in which players eventually cooperate with each other indefinitely. For example, for standard stage-game payoffs.[8] and arbitrarily patient players, the strategy profile given by Figure 3 constitutes an EXE:



Player 1: D D D C D C C C ...
Player 2: D C D D D C C C ...

Figure 3

On the other hand, the same Tit-for-Tat-like scheme cannot achieve cooperation in the case of repeated "Chicken".[9] I believe that this result uncovers an interesting difference between "Chicken" and the Prisoner's Dilemma: given that players test relevant threats along the equilibrium path, building

[8]

	C	D
C	3,3	0,4
D	4,0	1,1

[9] The stage-game payoffs are given by:

	D	H
D	3,3	1,4
H	4,1	0,0

up cooperation in repeated "Chicken" is more difficult than in the repeated Prisoner's Dilemma. Detailed discussion of this example is carried out in Spiegler (2001).

4 Conclusion

The concept of EXE captures a natural procedure of experimentation with sub-optimal behavior, namely testing threats that affect optimal response. The concept is based on two innovations - first, a formalization of the procedure in the context of repeated games and second, its incorporation in equilibrium behavior.

Of course, this is not the only conceivable type of experimentation. E.g., even when a player in the repeated Prisoner's Dilemma expects his opponent's strategy to be "always defect", he may occasionally experiment with cooperative behavior. This type of experimentation is ruled out by EXE. How can we justify the assumption that the players' experimentation activity in a repeated game is *exclusively* captured by EXE?

One possible answer draws on an analogy with classical statistics. When an "applied" classical statistician tests a multi-variable regression equation, "standard procedure" - the kind of procedure that is embedded in standard statistical software packages - involves running sample error tests (t-tests). Such a test examines the equation against a null hypothesis that sets one or more of the equation's parameters to zero. A wholly different procedure involves specification error tests, which examine the equation against the hypotheses that the equation omits certain relevant variables. These are two distinct procedures - the statistician sometimes carries out both of them, but often she carries out just the sample error tests. Ignoring the specification error testing procedures would be a fair approximation of the statistician's "normal" behavior.

Another answer draws on the justifiability rationale behind EXE. Players depart from optimizing behavior only if it helps them justify their behavior ex-post. The situations in which such departures are needed depend on the burden of proof criteria that face the player in the justification process. This idea is formalized in an earlier paper (Spiegler (1999)). The formal links between EXE and the idea of justifiability are explored in Spiegler (2001).

5 References

1. Aumann R. J. (1987). "Correlated Equilibrium as an Expression of Bayesian Rationality." *Econometrica* **55**, 1-18.

2. Osborne M. and A. Rubinstein (1994). A Course in Game Theory. MIT Press.

3. Rubinstein A. (1986). "Finite Automata Play the Repeated Prisoner's Dilemma." *Journal of Economic Theory* **39**, 83-96.

4. Rubinstein A. (1991). "Comments on the Interpretation of Game Theory." *Econometrica* **59**, 909-924.

5. Spiegler R. (1999). "Reason-Based Choice and Justifiability in Extensive Form Games". Foerder Institute Working Paper No. 19-99.

6. Spiegler (2001). "Procedurally Rational Experimentation in Infinite-Horizon Games." Mimeo.