The Illusion of Leadership: Misattribution of Cause in Coordination Games

Roberto Weber • Colin Camerer • Yuval Rottenstreich • Marc Knez
Department of Social and Decision Sciences, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213
Division of the Humanities and Social Sciences, 228–77, California Institute of Technology, Pasadena, California 91125
Graduate School of Business, University of Chicago, 1101 E. 58th Street, Chicago, Illinois 60637
Lexecon Strategy Group, Chicago, Illinois 60604
rweber@andrew.cmu.edu • camerer@hss.caltech.edu • yuval.rottenstreich@gsb.uchicago.edu • marc.knez@gsb.uchicago.edu

Abstract

This paper reports the results of experiments which examine attributions of leadership quality. Subjects played an abstract coordination game which is like many organizational problems. Previous research showed that when larger groups play the game, they rarely coordinate on the Pareto-optimal (efficient) outcome, but small groups almost always coordinate on the efficient outcome. After two or three periods of playing the game, one subject who was randomly selected from among the participants to be the "leader" for the experiment was instructed to make a speech exhorting others to choose the efficient action. Based on previous studies, we predicted that small groups would succeed in achieving efficiency but that large groups would fail. Based on social psychological studies of the fundamental attribution error, we predicted that the subjects would underestimate the strength of the situational effect (group size) and attribute cause to personal traits of the leaders instead—leaders would be credited for the success of the small groups, and blamed for the failure of the large groups. This hypothesis proved true: Subjects attributed differences in outcomes between conditions to differences in the effectiveness of leaders. In a second experiment, subjects voted to replace the leaders more frequently in the large-group condition (at a small cost to themselves), showing that misattributions of leadership ability also affect actual behavior by subjects. Previous research has demonstrated a tendency to credit or blame leaders for unusual performance. The difference in our study is that subjects should be blaming a structural condition—the size of the group —but they blame the leaders instead. Thus, our experiment is the first to establish a mistaken illusion of leadership.

(*Leadership*; *Coordination*; *Attribution Errors*; *Game Theory*; *Synergy*; *Mutualism*)

1. Introduction

Leadership seems to be important to the success or failure of every organization. As a result, corporations spend millions searching for people judged to be potentially effective leaders, and training and evaluating them.

There are basically two ways to judge leadership ability. One way is to rely on theories of what qualities good leaders have, and look for people with those qualities (see Stogdill 1948). For instance, we can look for a dynamic, well-spoken, confident, tall, attractive man, or a tough negotiator, or someone who is charismatic and friendly.

Another way to judge leadership ability is to use a person's past history of leadership success. For instance, we can evaluate a manager's leadership ability by the extent to which he increased profits at his previous company, or we can vote for the presidential candidate who, as governor, best improved his state's economy. Of course, these judgments require us to separate the effect of an individual's leadership ability from the difficulty of the situation in which they are asked to lead. Separating the two components requires an experiment in which we place this manager along with many others in the same position and observe how well he or she performs relative to the others. Such experiments rarely occur—except in the laboratory.

In this paper we describe experiments designed to see whether people mistakenly overattribute success and failure, which should be attributed to differences in situational difficulty, to differences in leadership ability. Subjects are randomly chosen to lead groups in two different conditions. In one condition success is relatively easy; in the other condition success is difficult to achieve. Based on social psychological studies of the "fundamental attribution error," we predicted that subjects would tend to mistakenly blame people—the leaders—for effects that are caused by the situational difference. If so, then people have an "illusion of leadership." Of course, if we observe an illusion it does not imply that there is no such thing as true leadership ability, or that people never know

Organization Science, © 2001 INFORMS
Vol. 12, No. 5, September–October 2001, pp. 582–598

1047-7039/01/1205/0582/$05.00
1526-5455 electronic ISSN

whether they have been led well or not. It simply establishes a tendency to adjust insufficiently for the difficulty of situational conditions in evaluating leaders.

2. Previous Research

Our paper weaves together three strands of previous research. One strand, from social psychology, studies how people make attributions of cause for outcomes they observe. Another strand, from game theory, provides us with a simple game that models some kinds of organizational decision making, and which provides conditions under which success or failure are likely to occur. The third strand is research on organizational behavior. Inspired by the psychological evidence of misattributions, this research asks whether people in organizations "romanticize" leadership and blame or credit their leaders too much.[1] Our experiments continue this research, but establish stronger findings for three reasons: Outcomes are clearly caused by a situational variable which subjects overlook when evaluating leaders; subjects participate in the situations rather than simply read about them or watch them; and leadership ratings are measured by whether subjects "fire" leaders (at a cost to themselves) as well as by rating scales. The importance of these three features will become apparent in our discussion of previous research, which we postpone until after our own results are presented.

2.1. Attribution

The tendency to overattribute the causes of behavior to personal traits rather than to aspects of a situation is called the "the fundamental attribution error" by social psychologists (see Ross and Nisbett 1991). Consider a study by Ross et al. (1977). These authors had subjects play a quiz game. One randomly selected subject, the questioner, generated 10 "challenging but not impossible questions" which a second subject, the contestant, tried to answer. A third subject, the observer, watched. At the end of the session the observers were asked to rate both the questioner's and contestant's general knowledge. The questioner's role advantage guaranteed that they would reveal some knowledge and could not seem stupid, which contestants might. Observers did not adjust for the impact of this aspect of the situation: They judged questioners to be much more knowledgeable than contestants. The robustness of the tendency to underweigh situational effects, which has been replicated in many experiments, is what has motivated psychologists to call it the fundamental attribution error.[2]

2.2. The Situational Variable: Group Size in Weak-Link Games

To establish an error in attributions of leadership ability, one needs to find a situational variable which is strongly related to likely outcomes, so that subjects may misattribute its effect to leaders. We found such a variable in experimental studies of coordination games.

We chose a simple "weak-link coordination game" which abstractly models many organizational situations (see Camerer and Knez 1997, and the conclusion below). These games were first studied experimentally by Van Huyck et al. (1990).[3] In our game, payoffs are a function both of a player's choice of a personal fee, or contribution, and the minimum personal fee chosen by any player. The payoff table is shown in Table 1. Each cell shows the payoff corresponding to a player's choice of a personal fee and the minimum personal fee chosen in that player's group (including the player's own choice).

The diagonal cells correspond to outcomes in which the player is choosing the same fee as the minimum. In a Nash equilibrium, everyone chooses the same fee and gets the same outcome. All fee levels are Nash equilibria. That is, if everyone thinks the minimum fee will be \$2, for example, then by choosing a \$2 fee they earn \$1.75. If they chose a smaller fee, like \$1, they would drag the minimum down to \$1 and earn less, \$1.50. If they chose the higher fee, \$3, they would earn even less, \$0.75. So choosing a \$2 fee is the best response to an expected minimum of \$2.

Notice that the equilibria are different, because those with higher personal fees also yield higher payoffs. The Pareto-dominant (or "efficient") outcome arises when all of the participants select the highest fee, \$3, and receive a net payoff of \$2. It is in the players' mutual interest to reach this outcome and the players almost certainly realize this.

However, the efficient outcome may not be easy to achieve. Choosing \$3 can yield the highest payoff, but it can also lead to low earnings. For instance, if just one of the other subjects playing the game selects \$0, then any subject playing \$3 loses \$2. Players are faced with strategic uncertainty. Simply being unsure about what others will do may lead different players to take different actions, and when groups are large the lowest personal fee may therefore be quite low.

Table 1 **The Payoff Matrix for Experiment 1 Weak-Link Game**

Player's Fee	Min = \$0	Min = \$1	Min = \$2	Min = \$3
\$0	\$1.00			
\$1	\$0.00	\$1.50		
\$2	−\$1.00	\$0.50	\$1.75	
\$3	−\$2.00	−\$0.50	\$0.75	\$2.00

Note that the weak-link game differs fundamentally from the Prisoner's Dilemma (PD) which is frequently studied by organizational researchers. In the PD every player prefers to contribute nothing, or "defect," regardless of what others do, and the result is an inefficient (Pareto-dominated) outcome. In the weak-link games, subjects prefer to reciprocate what others do: If others choose a low fee, they want to choose a low fee; but if others choose a high fee, it is in players' best interests to do so as well. Put more formally, in the weak-link game there is no dominant strategy (a strategy which is best regardless of what others do) as in PD.

The weak-link game is structurally similar to threshold or step-level public goods games (in which a public good is provided if enough subjects contribute), the "volunteer's dilemma" (Murnighan et al. 1993), and to infinitely repeated prisoners' dilemma games where trigger strategies create multiple Pareto-ranked equilibria. It is also an n-person, four-action version of the two-player "stag hunt" or "assurance" game.[4] In organizational settings, people play a weak-link game any time an organizational outcome is a sufficiently increasing function of the lowest-quality input provided by each person, and inputs are costly to people. For example, Knez and Simester (1997) discuss the interdependence present in airline departures which creates a weak-link coordination problem. For a plane to depart, several different tasks must be completed by different workers (loading passengers, baggage and food; preflight checks; fueling). The flight cannot depart until the last procedure is completed.[5]

Previous experiments with weak-link games have established clear regularities. Coordination on the efficient equilibrium has never been observed with large groups. Of the seven sessions initially conducted by Van Huyck et al. (1990) (VHBB) with groups of size 14 to 16, after the third period the minimum in *all* sessions was the lowest possible choice. For small groups ($n = 2$) coordination on the efficient equilibrium was much easier—it was reached in 12 of 14 (86%) of the groups studied (a result replicated by Knez and Camerer 1996). Table 2 summarizes the distribution of fifth-period minima in several different experiments, all using the VHBB game in which subjects choose integers from one to seven. Choosing seven is efficient.

The effect of group size could hardly be stronger. By placing subjects in a group of Size 2, we are almost assuring that they coordinate on the efficient outcome. By placing subjects in large groups of six or more, we are almost assuring that they will converge to the least efficient outcome. Thus, this game provides a simple situational variable—group size—which can easily be manipulated to strongly influence the outcomes.

Furthermore, there is indirect evidence that subjects don't realize how strong the group size effect is. Table 3 shows the distribution of first-period choices across various group sizes (from two to 14). The medians are underlined. If subjects realized that the minimum tends to be lower in a larger group, they should pick lower numbers in larger groups—since their goal is to match the minimum. But in fact, the distributions are remarkably similar across all group sizes in the first period. Subjects don't seem to be aware of the effect of group size.

The basic idea underlying all of our experiments is to randomly choose "leaders" to speak briefly to their group about what players in the group should choose. The prediction is that the leaders' speeches will be ineffective in large groups and unnecessary in small groups, because a short speech is unable to offset the strong causal influence of group size. Then players in large groups may mistakenly attribute their inability to achieve the best outcome to poor leadership rather than to situational misfortune; similarly, players in small groups may mistakenly attribute their ability to achieve the best outcome to good

Table 2 **Fifth Period Minimums (by %) in Various Seven-Action Weak-Link Studies (1 = Inefficient; 7 = Efficient)**

Minimum							Group		
1	2	3	4	5	6	7	Size	N	Source
9	0	0	0	0	0	91	2	28	VHBB 1990; Knez and Camerer 1996
37	15	15	11	0	4	18	3	60	Knez and Camerer 1994, 1996
80	10	10	0	0	0	0	6	114	Knez and Camerer 1994
100	0	0	0	0	0	0	9	18	Cachon and Camerer 1996
100	0	0	0	0	0	0	14–16	104	VHBB 1990

Table 3 **First-Period Choices (by %) in Various Seven-Action Weak Link Studies (1 = Inefficient; 7 = Efficient)**

Choice							Group		
1	2	3	4	5	6	7	Size	N	Source
17	2	2	6	16	0	*57*	2	28	VHBB 1990; Knez and Camerer 1996
7	4	9	15	*12*	2	45	3	60	Knez and Camerer 1994, 1996
14	8	13	*16*	3	3	37	6	114	Knez and Camerer 1994
0	11	28	*39*	5	0	17	9	18	Cachon and Camerer 1996
2	5	5	17	*32*	9	31	14–16	104	VHBB 1990

leadership rather than to situational good fortune. Such attributions are illusory perceptions of leadership.

3. Experiment 1: The Illusion of Leadership

A total of 79 subjects participated in the experiment. Of this total, 39 subjects were Stanford undergraduates recruited by means of ads posted around campus. They participated in the experiment on the Stanford campus, and received $7 for their time, in addition to any earnings from the game. The remaining subjects were high school students studying at Caltech during the summer. They were recruited by an announcement made in their class.[6] Subjects were run in groups of 10 (except for one group of nine) and were randomly assigned to one of two conditions. In the large group condition all of the subjects played the game together ($n = 39$). In the pairs condition ($n = 40$) subjects were randomly and anonymously paired with one of the other nine people in the room and played the game with this person only.

Subjects were instructed that they, along with every other person in their group, would each choose a "personal fee" of $0, $1, $2, or $3. The smallest personal fee would then determine the size of a reward to be paid to everyone in the group.[7]

Notice that a participant could assure herself of $1 by selecting a personal fee of $0 and thus determining the minimum. On the other hand, by choosing a personal fee of $3, a subject could earn $2 if all other participants chose a $3 fee, or could lose $2 if just one of the other players contributed $0.[8]

The experiment consisted of eight periods (eight plays of the game). At the beginning of every period, each subject circled their choice of personal fee from all of the alternatives on their record sheet. The experimenter then proceeded around the room and recorded the choices made by each subject. After recording all of the choices, the experimenter wrote the relevant information on a board at the front of the room. For large groups, the experimenter announced and wrote the smallest personal fee and resulting reward, while for pairs, the experimenter repeated this process five times, once for each pair.[9]

Lastly, before the first period the leader was selected by having the N subjects in a session draw balls from a bag containing N-1 white balls and one orange ball. The participant who drew the orange ball was designated as the leader for the entire experiment. Hence, the leader in the pairs condition addressed several pairs at once. The leader was told that he or she would, after the second period, make one address in order to "organize" and "prepare" the players for the remaining rounds, and was given the following handout:[10]

> Please deliver the message below to the other participants. You'll be more convincing if you don't read. You don't have to quote the message word for word. Paraphrasing is allowed.
> **Leader's Message**: We need to coordinate here. Obviously, we all do best if everybody chooses a personal fee of $3. That makes the reward $5, and gives each of us earnings of two bucks per round. Let's not be dumb here.

Subjects completed a questionnaire after the second period of play, but before the leader's speech. During this time, the experimenter checked with the leader to make sure that he or she was ready to speak. The leader then spoke briefly, usually for less than 30 seconds, at which point subjects completed a second questionnaire. At this point, subjects played the remaining six periods and then completed a final questionnaire.

Note that what we are calling leadership is a simple address by one randomly selected group member, which does not contain many of the elements usually associated with leadership. We want to see whether an illusion of leadership occurs in a simple situation before adding more complicated features.[11] Moreover, if we used a more elaborate form of leadership, selected subjects would probably appear more like actual leaders, and hence more deserving of credit and blame. Thus, we see our experiment as establishing a lower bound on the amount of misattribution which occurs.

3.1. Results

The choices of personal fees across periods are reported in Table 4.[12] For rounds before and after the leaders' speeches, the distributions of pooled personal fees for the large group ($n = 78$ before, $n = 234$ after) and for the pairs ($n = 80$ before, $n = 240$ after) are reported. In addition, we report the results of a one-tailed Kolmogorov-Smirnov (KS) test of the difference between the distributions. The fees are somewhat different in Rounds 1 and 2, but that difference is not highly significant ($p < 0.1$). This is consistent with the idea that subjects largely fail to anticipate the group-size effects (too many subjects choose high fees in the large groups in the first two rounds).

Table 5 presents the answers to the questions players were asked.[13] All of the questions elicited a rating on a nine-point scale.

Notice first that immediately after the leaders speak they are judged about the same whether they spoke to a large group or to small groups; the median judgment of overall leadership ability is a six in either case and the means are very close (5.88 for pairs and 5.80 for large groups) and insignificantly different. Immediately after the speeches, subjects thought these were both equally good leaders.

Table 4 Distribution of Fees Before and After Leader's Speech in Experiment 1

	Group Size	\$0	\$1	\$2	\$3	Mean	Test for Difference KS(1)
Rounds 1–2 (prespeech)	9–10	25%	24%	20%	32%	\$1.579	$\chi_2^2 = 5.518, p < 0.1$
	2	5%	24%	26%	45%	\$2.105	
Rounds 3–8 (postspeech)	9–10	47%	4%	0%	49%	\$1.500	$\chi_2^2 = 69.579, p < 0.001$
	2	6%	6%	6%	83%	\$2.658	

Table 5 Questions and Responses for Experiment 1

	Size	Median	Mean	$\hat{\sigma}$
Questions asked immediately after Round 2				
1. In your opinion how good or bad is the judgment of the other participants here today?	9–10	4.5	3.74	1.81***
(1 = Very Poor; 9 = Excellent)	2	5.0	5.94	1.81
Questions asked immediately after leader speaks				
1. How well has the leader prepared the participants for the next six rounds?	9–10	6.5	6.62	2.05 N.S.
(1 = Extremely Poorly; 9 = Extremely Well)	2	7.0	6.83	1.56
2. Please rate the leader's overall leadership ability	9–10	6.0	5.88	1.89 N.S.
(1 = Extremely Poor; 9 = Extremely Good)	2	6.0	5.80	1.23
Questions asked after Round 8				
1. In retrospect, how good or bad is the judgment of the other participants here today?	9–10	4.0	4.21	2.48***
(1 = Very Poor; 9 = Excellent)	2	7.0	6.80	1.88
2. In retrospect, how well did the leader prepare the participants for the last six rounds?	9–10	4.0	4.68	2.75***
(1 = Extremely Poorly; 9 = Extremely Well)	2	7.0	6.74	1.77
3. Please rate the leader's overall leadership ability.	9–10	4.0	4.53	2.31***
(1 = Extremely Poor; 9 = Average)	2	7.0	6.17	1.71
4. Consider the leader's task of organizing and preparing the participants, was it:	9–10	6.0	5.91	1.81**
(1 = Extremely Difficult; 9 = Extremely Easy)	2	7.5	6.74	2.05

*** – $p < 0.01$; ** – $p < 0.05$; * – $p < 0.1$; N.S. – Not Significant

Following the leader's address, however, the outcomes differed considerably between the two groups. As Table 4 indicates, in postspeech rounds subjects playing in pairs select the highest (\$3) and lowest (\$0) personal fees 83% and 6% of the time, compared to 49% and 47% in large groups. As a result, while a minimum of \$3 was achieved in 79 percent of the pairs' outcomes, a minimum of \$0 occurred in 75 percent of the trials in large groups.

After all eight rounds, leaders in the pairs condition were judged to be effective; the mean rating of their "overall leadership ability" (Question 3) was 6.17, compared to 5.80 immediately after their speech. In contrast, leaders in the large groups were judged relatively ineffective; the mean rating of their overall leadership ability was 4.53, compared to 5.88. A comparison of the means (6.17 for pairs and 4.53 for large groups) is highly significant.[14]

Thus, subjects experienced an illusion of leadership. They mistakenly attributed good outcomes to good leadership ability and bad outcomes to poor leadership ability when in fact these outcomes occurred because of the nature of the situation in which the subjects were placed.[15]

Illusions are not limited to vague generalizations about overall leadership ability. Subjects were also asked "How well did the leader prepare the participants for the last six rounds?" This judgment requires only a reading of the leader's performance and no extrapolation to a general ability, but even here an illusion arose. Just after the speech, leaders in the two conditions were seen as equally effective in preparing the participants; pairs leaders received median and mean ratings of 7 and 6.83, while large group leaders received ratings of 6.5 and 6.62 (which are insignificantly different). After the last six rounds had been played, pairs leaders received a median and mean rating of 7 and 6.74 (very close to their postspeech ratings), while large groups leaders were marked down to a median of 4 and a mean of 4.68 (significantly different from the pairs ratings at $p < 0.01$ in both tests).

Judgments about other players and about situational difficulty also exhibited interesting effects. Subjects were asked "how good or bad is the judgment of the other participants here today?" (1 is "extremely poor" and 9 is "excellent"). Subjects mistakenly blame others for the bad outcomes caused by group size: Subjects in the pairs condition gave mean ratings of 5.94 after two rounds and 6.80 after all eight rounds, significantly higher than the corresponding ratings of 3.74 and 4.21 in the large groups condition ($t_{67} = 5.048$ for round two and $t_{67} = 4.88$ for round eight, both $p < 0.001$). While we have stressed the issue of leadership in organizations, these results indicate that morale and blame of fellow workers may be equally prone to illusions and misattributions.

Subjects also recognized the difference in the leadability of the situations. When asked about the difficulty of "the task of organizing and preparing the participants" (Question 4), subjects in the large groups respond that the leader's task was significantly more difficult ($t_{65} = 1.76$, $p < 0.05$ by a one-tailed test). This shows that while subjects realize the situational effect, they fail to adjust sufficiently for it when judging leaders (and other subjects in their groups).

The situational difficulty ratings are important because some social psychologists have argued that people generally attribute more cause to forces which are focal or perceptually salient (central, visually available, or proximate in time or space; see Taylor and Fiske 1975). Cause is also readily attributed to forces which are "mutable," or easy to imagine having occurred otherwise (Kahneman and Miller 1986). In this interpretation, situations are underweighted compared to personal traits, simply because actions of people tend to be more salient or mutable than situational factors.[16]

In our experiment, leaders are perceptually salient because they stand at the front of the room and make a speech. Their speech is mutable because it is easy for subjects to imagine a different speech. The mutability perspective predicts that if subjects were made more aware of the situational difference, the attribution to leaders would have been weakened. This could be an interesting topic for further research. However, the fact that subjects did rate the large-group situation as less leadable means they were aware of possible situational differences, and still attributed cause to leaders.

4. Experiment 2: Costly Voting as a Measure of Leadership Ratings

In Experiment 1, subjects rated the leaders on a nine-point scale. It is also important to see whether behavior with costly consequences to subjects reveals similar attributions of leadership. Experiment 2 does so by asking subjects, after eight rounds of play, to cast a costly vote to determine whether or not to replace the leader for an additional set of rounds.

Experiment 2 consisted of two parts. The first part was an exact replication of Experiment 1. Subjects played the game described above for eight rounds in either large groups or pairs, leaders were randomly selected at the beginning of the experiment, and the same questions were used to elicit ratings of leadership quality. In order to exactly replicate Experiment 1, subjects were not informed during the first part of Experiment 2 that there would be a second part to the experiment.

Once the first part was completed, subjects were told that they would now play an additional four rounds of the game above in the same groups as before, and that they could keep the leader or randomly select a new one. The process was described as follows:

> You now have the opportunity to vote to either have a new leader or keep your current session leader. Every participant in the room, with the exception of the current leader, will cast a secret vote to either "Keep the Leader" or "Have a New Leader". All of these votes will be counted, and if more participants vote to "Have a New Leader" then a *new* leader will be randomly selected from among the participants in the room, and everyone who voted to "Have a New Leader" will be charged 25 cents. In the case of a tie or if there are more votes to "Keep the Leader," the current leader will *not* be replaced and no participants will pay the charge.

Notice that there is a small cost associated with replacing the leader.[17] Since the vote determines the leader for the subsequent set of rounds, subjects now have a monetary incentive to respond correctly when voting whether or not to replace the leader. If they believe that the current leader is truly effective (that the leader will induce everyone to coordinate on the efficient equilibrium), then they should want to keep this leader for the remaining rounds, while if they believe that the leader is less likely to produce favorable results, then they should vote to replace him or her (which could earn them substantially more money). This differs from the method used in Experiment 1, where subjects' approval of the leader was measured solely through question responses.

At the conclusion of the vote, the outcome was announced to all the subjects. If more participants voted to "Have a New Leader," then a new leader was randomly selected; otherwise, the leader remained the same. The leader then gave a short speech to the group similar to the one before. After the speech, four rounds were played in the same manner as before.

For Experiment 2, subjects were undergraduates at the California Institute of Technology. A total of 81 subjects were used (41 in large groups and 40 in pairs).

4.1. Results

The first part of Experiment 2 consisted of an exact replication of Experiment 1. The results of this replication are presented in Tables 6 and 7, which provide an aggregate summary of the personal fees and questionnaire responses, respectively. These tables correspond to Tables 4 and 5, which present the same results for Experiment 1.

As Table 6 shows, the average personal fee again falls between early and late rounds for the large groups and rises for the pairs. However, the personal fee choices in the first two (prespeech) periods differ across conditions more than in Experiment 1 ($p < 0.001$). The important question is whether the difference in outcomes increases after the address by the leaders (rounds 3–8), which it does. Indeed, the fact that the small and large groups already behaved differently before the leaders' speeches should undermine the tendency of subjects to misattribute cause to the leaders.

Table 7 provides responses to the survey questions asked of subjects during the first part of the experiment. The ratings replicate Experiment 1 quite well. The mean responses to the question about "overall leadership ability" are close together immediately after the leaders' speech (4.92 for pairs and 4.97 for large groups, insignificantly different). At the end of eight rounds of play, however, the responses to the same question are much further apart (means of 4.62 for pairs and 3.82 for large groups) and significantly different ($t_{73} = 1.828$; $p < 0.05$).

Table 6 Distribution of Fees Before and After Leader's Speech in the First Part of Experiment 2

	Group Size	\$0	\$1	\$2	\$3	Mean	Test for Difference KS(1)
Rounds 1–2 (pre-speech)	13–15	33%	21%	17%	29%	\$1.472	$\chi^2_2 = 25.249$, $p < 0.001$
	2	10%	9%	13%	69%	\$2.400	
Rounds 3–8 (post-speech)	13–15	52%	14%	8%	26%	\$1.065	$\chi^2_2 = 193.724$, $p < 0.001$
	2	10%	0%	1%	89%	\$2.679	

Table 7 Questions and Responses for the First Part of Experiment 2

	Size	Median	Mean	$\hat{\sigma}$
Questions asked immediately after Round 2				
1. In your opinion how good or bad is the judgment of the other participants here today?	13–15	4.5	3.92	2.05***
(1 = Very Poor; 9 = Excellent)	2	6.0	5.84	2.53
Questions asked immediately after leader speaks				
1. How well has the leader prepared the participants for the next six rounds?	13–15	5.5	5.55	2.51 N.S.
(1 = Extremely Poorly; 9 = Extremely Well)	2	7.0	6.30	1.97
2. Please rate the leader's overall leadership ability	13–15	5.0	4.97	2.54 N.S.
(1 = Extremely Poor; 9 = Extremely Good)	2	5.0	4.92	1.88
Questions asked after Round 8				
1. In retrospect, how good or bad is the judgment of the other participants here today?	13–15	3.0	3.42	1.80***
(1 = Very Poor; 9 = Excellent)	2	7.0	6.35	2.23
2. In retrospect, how well did the leader prepare the participants for the last six rounds?	13–15	3.0	3.66	2.08***
(1 = Extremely Poorly; 9 = Extremely Well)	2	6.0	5.81	2.22
3. Please rate the leader's overall leadership ability.	13–15	3.0	3.82	2.05**
(1 = Extremely Poor; 9 = Average)	2	5.0	4.62	1.75
4. Consider the leader's task of organizing and preparing the participants, was it:	13–15	5.0	4.59	2.05***
(1 = Extremely Difficult; 9 = Extremely Easy)	2	7.0	7.22	1.65

*** – $p < 0.01$; ** – $p < 0.05$; * – $p < 0.1$; N.S. – Not Significant

The final question in Table 7, concerning the ease or difficulty of the leader's task, provides some insight into why the differences in final leadership ratings are not as large as in Experiment 1. The difference across conditions is much larger in Experiment 2 (7.22 for pairs vs. 4.59 for large groups, significant at the $p < 0.001$ level by a t-test ($t_{73} = 6.102$)) than it was in Experiment 1 (6.74 vs. 5.91). Thus, the subjects in Experiment 2 were simply more aware of the group-size effect.[18]

The main point of Experiment 2 is to test whether or not the results from Experiment 1 replicate when subjects are asked to vote on replacing the leader. As hypothesized, more subjects vote to replace the leader in the large groups (32%, $n = 38$) than in the pairs (16%, $n = 37$). While small, this difference is significant at $p = 0.06$ ($t_{73} = 1.565$, one-tailed).

To test the relation between the rating a subject gave a leader and that subject's subsequent vote, we conducted a logit regression with a subject's vote ("Have a New Leader" (0) or "Keep the Leader" (1)) as the dependent variable. For independent variables, we used treatment (pairs (0) or large groups (1)), the final round minimum in that subject's group, and the subject's responses to three of the questions asked after Round 8 (Questions 1, 3, and 4 at the bottom of Table 7). The only significant coefficient obtained was for the question regarding "overall leadership ability" (coefficient = 0.452, t-statistic = 2.50). This means that the *only* significant determinant of whether or not a subject votes to replace the leader is that subject's rating of leadership quality.[19] This is striking because it means that the outcomes in a subject's group do not affect her vote directly: Subjects do not vote to replace leaders in low-outcome groups per se, they vote to replace leaders to whom they gave low ratings.

The Experiment 2 results replicate the Experiment 1 results. The fact that the effect persists even when subjects are more strongly aware of the situational effect on outcomes gives some indication of its strength. More importantly, the results of the vote indicate that subjects are willing to act on these attributions; the previous results are not merely an artifact of using rating scales to measure perceptions of leadership ability. In the next section we address the question of whether the results of Experiment 1 can be replicated in a more contextually rich situation.

5. Experiment 3: Adding Realism

One criticism of the results of Experiments 1 and 2 is that instructing subjects to simply choose "personal fees" so that they can receive a "reward" does not create a realistic situation. Attributions of leadership quality might not arise in a more contextually rich setting. To address this concern, we conducted an experiment which differed from Experiment 1 in only one way—instructions presented subjects with a more realistic and familiar task. The instructions read:

> In this experiment, you are one of N members of a project team that is responsible for producing a series of reports. Each report that the team prepares consists of N sections, where each member of the team is responsible for contributing one of the sections. A report is considered complete only after all members of the team contribute their sections. Your team will be responsible for producing a total of eight reports. Until a particular report is finished, no member of the team can work on his or her section of the next report.
>
> You earn money based on how rapidly each report is completed. Each report is due in 4 weeks, however, every team member receives a bonus if the team completes the report in less than four weeks. There are three possible early completion times: 1 week, 2 weeks, or 3 weeks ahead of schedule. Hence, as a team member you must decide whether to contribute your section of a report during Week 1, Week 2, Week 3, or Week 4. The earlier a report is completed, the larger the bonus.

Subjects were given the "personal contribution time costs" and "completion time rewards" associated with each of the four weeks, and these costs and rewards were identical to the ones used in Experiments 1 and 2. Notice that earlier completion times correspond to *higher* personal fees. Efficiency is reached if all subjects choose Week 1 (three weeks early).

The experiment was otherwise conducted identically to Experiment 1. Two sessions were conducted for each condition ($n = 16$ for pairs and $n = 17$ for large groups).[20] Subjects were University of Chicago undergraduates.

5.1. Results

The choices are shown in Table 8.[21] These results closely replicate those of Experiment 1. There is no significant difference in choices between the two conditions for the prespeech periods. After the leaders' speeches, however, subjects in the pairs condition coordinate on the efficient equilibrium while subjects in large groups do not.

The questionnaire responses also replicate the results of Experiment 1. Table 9 presents the mean and median responses by condition (compare with Tables 5 and 7). Ratings of "overall leadership quality" immediately after the speeches indicate that leaders in the pairs condition were rated slightly more favorably (5.64 for pairs vs. 4.67 for large groups). However, this difference is not significant at any reasonable levels ($t_{27} = 1.178$). Ratings for the leaders in the pairs condition rise (6.57) while the ratings for leaders in large groups fall (3.13) by the final

Table 8 Distribution of Choices Before and After Leader's Speech in Experiment 3

	Group Size	4($0)	3($1)	2($2)	1($3)	Mean	Test for Difference KS(1)
Rounds 1–2 (prespeech)	8–9	6%	24%	38%	32%	2.029	$\chi_2^2 = 2.845$, N.S.
	2	6%	6%	34%	53%	1.656	
Rounds 3–8 (postspeech)	8–9	56%	15%	5%	25%	3.020	$\chi_2^2 = 83.767$, $p < 0.001$
	2	9%	1%	0%	90%	1.302	

Table 9 Questions and Responses for Experiment 3

	Size	Median	Mean	$\hat{\sigma}$
Questions asked immediately after Round 2				
1. In your opinion how good or bad is the judgment of the other participants here today?	8–9	4.0	3.67	2.16***
(1 = Very Poor; 9 = Excellent)	2	5.5	5.14	1.70
Questions asked immediately after leader speaks				
1. How well has the leader prepared the participants for the next six rounds?	8–9	6.0	5.73	2.28 N.S.
(1 = Extremely Poorly; 9 = Extremely Well)	2	7.0	6.36	2.50
2. Please rate the leader's overall leadership ability	8–9	5.0	4.67	1.76 N.S.
(1 = Extremely Poor; 9 = Extremely Good)	2	5.5	5.64	2.62
Questions asked after Round 8				
1. In retrospect, how good or bad is the judgment of the other participants here today?	8–9	2.0	2.86	1.83***
(1 = Very Poor; 9 = Excellent)	2	7.0	6.50	2.03
2. In retrospect, how well did the leader prepare the participants for the last six rounds?	8–9	3.0	3.13	1.55***
(1 = Extremely Poorly; 9 = Extremely Well)	2	8.0	7.43	2.10
3. Please rate the leader's overall leadership ability.	8–9	3.0	3.13	1.46***
(1 = Extremely Poor; 9 = Average)	2	7.5	6.57	2.38
4. Consider the leader's task of organizing and preparing the participants, was it:	8–9	6.0	5.80	2.34***
(1 = Extremely Difficult; 9 = Extremely Easy)	2	8.0	8.00	1.18

*** – $p < 0.01$; ** – $p < 0.05$; * – $p < 0.1$; N.S. – Not Significant

period, and this difference is highly significant ($t_{27} = 4.696$, $p < 0.001$). Thus, the magnitude and the significance of these results replicate Experiment 1.

It is also interesting to note that subjects were again strongly aware of a group-size effect (as in Experiment 2). Subjects in large groups rate the leader's job as significantly more difficult than do subjects in the pairs condition. It may be that the familiar context of groups facing a deadline, and waiting for the slowest member, cues subjects to the relative difficulty of coordinating larger groups. But they blame and credit leaders anyway, despite their awareness of the situational difficulty.

Experiment 3 replicates the effect observed in Experiments 1 and 2. More importantly, it shows that a more realistic context does not weaken the misattributions to leadership quality. In fact, the change in leadership ratings is *larger* in magnitude when the context is provided. This supports a claim we made in earlier drafts of this paper, that the abstract context in Experiments 1 and 2 is, if anything, likely to understate the extent of misattribution.

6. Previous Organizational Research on Leadership Illusions

There is a widespread belief among organizational researchers that the illusion of leadership had been well established by previous research. This belief is wrong (and leads one to undervalue our contribution). Previous research suggested the possibility of such an illusion and reported data consistent with it. But some papers reported contradictory results and all others left open the possibility that an attribution of leadership ability is not a mistaken attribution, in a way we make clear below.

Our experiments establish that attribution is a mistake by using a situational variable which causes outcomes, and showing that subjects misattribute that cause to behavior by leaders even though they observe the situational variable. Our experiments also pushed further by using a "live" participatory activity in which subjects observe everything the leader does (only one previous leadership experiment did so) and, in our Experiment 2, by measuring attributions through subsequent voting behavior rather than simply rating scales (which no previous study has done).

The general argument made in previous research is that leadership is "romanticized" (Meindl et al. 1985, Meindl and Ehrlich 1987) because perceptions of the importance of leadership are exaggerated. For example, Meindl (1995 p. 330) writes

> The romance of leadership notion refers to the prominence of leaders and leadership in the way organizational actors and observers address organizational issues and problems, revealing a potential 'bias' or 'false assumption-making' regarding the relative importance of leadership factors to the functioning of groups and organizations.

The psychological underpinning of this view is that perceptions of leadership ability are attribution errors.[22]

One part of the argument that leadership perceptions are exaggerated is that the true effect of different leaders on performance is small. For example, Lieberson and O'Connor (1972) found that CEO identity was only weakly correlated with profitability of firms. Salancik and Pfeffer (1977) found that the identity of city mayors was only weakly correlated with city budgets (though the correlation was larger for discretionary funds not heavily influenced by interest groups). While these studies do suggest that the effects of leadership (or at least leader identity) are small,[23] they do not match up measured effects with perceived effects, so they do not establish that the perceived contribution of leaders is overestimated.

Another part of the argument is that performance tends to be attributed to leadership skill. Several experimental studies establish these attributions. Most studies use the following "performance cue paradigm": Subjects watch a videotape, listen to an audiotape, or read a vignette describing a group or firm's activity, including a leader's behavior. In high (low) performance-cue conditions, subjects are told that the group or firm performed well (or poorly) and asked to rate specific qualities or frequencies of behaviors by the leader and other group members. The crucial feature of this design is that the behavior on the tape or vignette is held constant while the performance cue varies. The typical finding is that subjects rate the incidence of effective leader behavior[24] as more frequent, or overall leadership skill as greater, when performance is better. Studies which use this paradigm and establish the attribution result include Mitchell et al. (1977), Lord et al. (1978), and Staw and Ross (1980).

Two studies departed from this experimental paradigm. Meindl et al. (1985) use the performance-cue paradigm but demonstrate a U-shaped relation between performance cues and ratings of the strength of leadership as a causal force—unusually good and bad performance is more strongly attributed to leadership. Their subjects did not rate leadership ability or behavior frequencies.[25] Staw (1975) used the performance-cue paradigm to measure how performance cues affect perceptions of group activity (like quality of communication, task conflict, and cohesiveness) but did not include leader behavior.

A problem with the performance-cue paradigm is that the conditions under which performance cues should affect ratings have not been clearly established. If there are aspects of leader behavior which are not clearly observed by subjects, and those unobserved aspects are related to performance, then subjects *should* use performance cues as indicators of unobserved leader behavior, and should rate leaders who performed well more highly (cf. Baron and Hershey 1988). To grasp our point, imagine an experiment in which medical students watch a videotape of portions of a medical operation, are told the outcome, then rate the quality of the operating surgeon. If the subjects realize there are small details of the operation they cannot see which affect its outcome, then it makes sense for the news that the patient died to affect their rating of the surgeon. The fact that the patient died supplies some information about what happened during the operation that they did not see.[26]

A precise way to make this point is through a model of leadership ratings and the effect of performance cues. Formally, let

$$L^{obs} = L + \tilde{L}, \tag{1}$$

where L represents the leadership ability of a given person. In a given leadership situation or role, this leadership ability is not observed, but instead observers see L^{obs} which is equal to leadership ability plus noise.[27] In a typical performance-cue experiment, subjects are trying to judge leadership ability but have only a sample of taped or written behavior to go on.

Assume that performance is determined by the following linear model:

$$P = \alpha_p + \beta_L L + \beta_S S + \tilde{\varepsilon}, \tag{2}$$

where S represents the situational variables which affect performance and $\tilde{\varepsilon}$ is a stochastic error term.[28]

From the two equations above, we can see that a rational observer who is unaware of the values of $\tilde{L}$ and $\tilde{\varepsilon}$ (but who is aware of the model) has two unbiased estimates of L:

$$\hat{L}_1 = L^{obs} \tag{3}$$

$$\hat{L}_2 = \frac{P - \alpha_P - \beta_S S}{\beta_L}. \tag{4}$$

Therefore, observers can use all their information to derive an estimate of L by performing a mental "regression" and assigning weight to both estimates:

$$L^{rated} = \hat{\gamma} L^{obs} + \hat{\beta}\left(\frac{P - \alpha_P - \beta_S S}{\beta_L}\right). \tag{5}$$

The measure L^{rated} is the rater's attempt to get as close to L as possible, using a weighted combination of the preperformance impression of ability, L^{obs}, and the situation-adjusted performance cue.

In evaluating leadership, we assume raters are facing a penalty such that they want to minimize the variance of the difference between their rating and the true quality of leadership (such as with a quadratic loss function). Equivalently, a conscientious rater is trying to choose weights $\hat{\gamma}$ and $\hat{\beta}$ to

$$\min_{\hat{\gamma},\hat{\beta}} Var[L^{rated} - L]$$

where $Var[X]$ denotes the variance of X.

Working through the solution to this minimization problem and simplifying gives (see appendix)

$$\hat{\beta} = \frac{1}{1 + \dfrac{\sigma_{\tilde{\varepsilon}}^2}{\beta_L^2\, \sigma_{\tilde{L}}^2} + \dfrac{\sigma_{\tilde{\varepsilon}}^2}{\beta_L^2\, Var[L]}}. \tag{6}$$

Equation (6) gives a way to specify conditions under which a rater should put some weight on the performance cue if she is trying to rate leadership ability as accurately as possible. The performance cues *should* matter more (i.e., $\hat{\beta}$ should increase) as leadership ability has an increased effect on performance (β_L increases) and as the noise component of observed leadership becomes greater ($\sigma_{\tilde{L}}^2$ increases). In addition, as performance becomes a less noisy measure of the effects of leadership and the situation ($\sigma_{\tilde{\varepsilon}}^2$ increases), then subjects should put more weight on the performance cues ($\hat{\beta}$ should increase). Thus, the fact that performance cues affect leadership ratings in the studies does *not* establish a mistaken attribution.

Furthermore, in the standard performance-cue paradigm, subjects try to rate leadership from a tape or vignette. In this paradigm, there will always be some unobservable information about leadership ability. Then $\tilde{L}$ has some variance and the performance cue should receive some weight. Our model predicts that as subjects' opportunity to observe the leader more closely and completely increases, $\sigma_{\tilde{L}}^2$ will fall and the weight they should assign to the performance cue, $\hat{\beta}$, will fall. For example, in participatory experiments the subjects observe everything the leader does and says, and also see the other subjects seeing that, which should reduce the variance of $\tilde{L}$ and lower $\hat{\beta}$ compared to the vignette experiments.

The findings of Mitchell et al. (1977) support this hypothesis. They report two studies in which subjects listened to or watched tapes. In both studies they find significant effects of the performance cue on subjects' ratings of two LBDQ measures of leadership skill, "consideration" and "initiation of structure." In a third study subjects actually participated in a group with a (confederate) leader to solve a business problem. In that study there were *no* significant effects on the two LBDQ measures (i.e., $\hat{\beta}$ was not significantly different from zero).[29]

Equation (6) shows that the existence of some unobserved component of leadership skill which affects performance should lead subjects to take performance into account when rating leadership. This means that we cannot conclude from the previous experiments that subjects are making mistaken attributions of performance to leaders. On the other hand, we cannot determine whether they are weighting performance optimally. They should weight performance positively, but they may be weighting it too highly.[30]

Because the subjects in our experiments observe everything the leader does, there is little unobserved leadership skill $\tilde{L}$ which affects performance in the game ($\tilde{L}$ is as small as we can make it experimentally). Therefore, in this case they should not use the performance cue (the outcomes in the game) to judge leadership skill. The fact that they do use the outcomes as a cue means their attributions are a mistake. Indeed, there is *no* previously-published experimental study in which, as in ours, subjects participated in a task with a leader they could observe and exhibited a significant performance-cue sensitivity in leader ratings. Thus, our experiments are the first to clearly establish a *mistaken* illusion of leadership, by using a situational variable which is causal and immersing subjects in a participatory activity.[31]

7. Conclusion

Our basic finding is that attributions of leadership ability depend on situational factors that subjects underappreciate. In Experiment 1, leaders in large and small groups make speeches which, immediately after the speeches, are rated equally highly. Then the large groups go on to fail, while the small groups succeed. Afterwards, the large-group leaders are considered good leaders and the small-group leaders are considered bad leaders. Experiment 2 shows that subjects are willing to act on these ratings by voting (at a cost) to "fire" leaders they give low ratings to—before continuing to play. Experiment 3 replicates the results even more strongly in a familiar context (group project with a time deadline).[32]

Figure 1 and Table 10 present the aggregated primary misattribution result. Figure 1 shows the average choice of personal fee or contribution time by condition for all three experiments. This graph provides evidence of the strength of the situational variable in our experiments. In all three experiments, the pairs converge on the efficient equilibrium while the large groups move toward the inefficient one. Note also that the choices are remarkably similar across conditions and experiments, both in the first period and in the third period, immediately after the leaders speak.

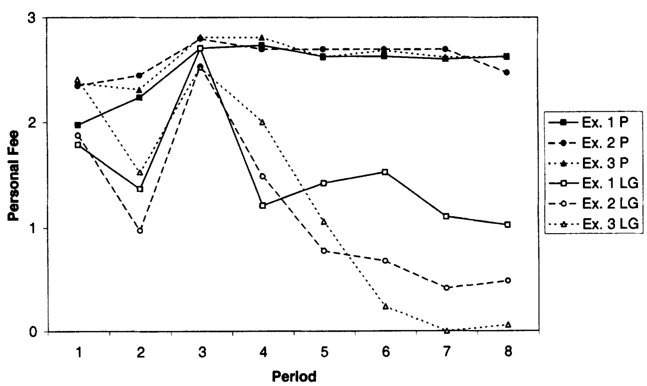
Table 10 shows the change in ratings of overall leadership quality (final rating minus initial rating) by condition for each experiment and for the aggregate data. Note that the ratings always fall in the large groups, that they rise in the pairs in Experiments 1 and 3, and fall slightly in the pairs in Experiment 2. The aggregate data reveal a strong illusion of leadership.[33]

It is also interesting to note in Table 10 that there appears to be an asymmetry in the credit and blame assigned

Table 10 Mean Change in Leadership Ratings by Condition and Experiment (Final Ratings—Initial Ratings)

Condition	Experiment 1	Experiment 2	Experiment 3	Aggregate
Pairs	0.37	−0.30	0.93	0.17
Large Groups	−1.35	−1.14	−1.53	−1.29

Figure 1 Average Personal Fee Choices by Experiment and Condition



to the leaders between conditions. Notice that in all three experiments, the magnitude of the average change is larger for the large groups than for the pairs. Aggregating the data, the leaders in the pairs condition are given an average increase in rating of 0.17 while the ratings for large group leaders fall by 1.29. This difference in magnitude is significant at the $p < 0.001$ level ($t_{170} = 3.989$). This asymmetry is consistent with the idea that subjects are more willing to credit themselves for good outcomes and blame others (the leaders) for poor outcomes.

In addition, ratings of the difficulty of the two situations are slightly different in Experiment 1 and more significantly different in Experiments 2 and 3. These differences show that subjects have *some* awareness that situations matter, but rate the two leaders differently despite that awareness.

While our primary focus is on leadership attributions, the coordination games could be used as an experimental paradigm to study other organizational phenomena. An economic process benefits from coordination, and hence can be modelled as a coordination game, if the optimal choice for one agent depends on what another agent does (so there are multiple equilibria). The crucial feature of coordination, mutual interdependence, is called "strategic complementarity" in game theory and industrial organization, mutualism in population ecology, and synergy or scope economy in business strategy literature. The fact that coordination is a central problem of organizing has been recognized repeatedly (e.g., Thompson 1967, Simon 1991, Milgrom and Roberts 1992), but the formal tools of game theory, and experiments like ours, are rarely used by organizational researchers.

Some recent papers show the ubiquity and subtlety of coordination in organizations. Kogut and Kulatilaka (1997) highlight the importance of choosing bundles of activities, when productivity of one bundle depends on the levels of other activities. Tushman and Murmann (1997) discuss changes in core and peripheral elements of technology as driving forces in innovation. Core elements are those which, when changed, render other technological features or elements of organizational structure obsolete or inefficient; changing core elements therefore requires coordination with other changes. (Peripheral elements, in contrast, can be changed individually and therefore do not require coordination.)

Winter and Szulanksi (1997) describe how a bank successfully replicates its business processes in ailing bank branches. An important element of the process is that the bank switches processes completely, from one system to another, during a one-day replacement period on a day announced well in advance. Switching so quickly incurs certain costs, but avoids the coordination failure which would occur if various parts of the system were switched at different points in time (like in the deadline version of the weak-link game in our Experiment 3). A related issue is why the bank cannot simply sell information about their process to the bank units, rather than painstakingly implementing it themselves. The answer is that while information can be reproduced cheaply (Xeroxing documents), understanding what it means requires a specialized language and know-how. In general, transferring information requires speaker and listener to coordinate on a common language (precise definitions of terms, agreement on implicit background assumptions, what rules can be violated and which are taboo, and so forth).[34] What the bank is implementing when it transfers business processes is not a plain language written in documents—which is easily "transferred"—but a detailed system of teaching employees how to communicate with each other in a new language, which is a kind of coordination problem.

7.1. Generalizing Our Findings

There are several kinds of organizational situations in which an illusion of leadership, like the one reported here, might arise. For example, Presidents who serve when both houses of Congress are controlled by the opposition surely operate in a less leadable situation than those with their party in Congress; do they get rated as worse Presidents, as a result? Is the head coach of a star-studded team which wins the championship recognized as better than the coach of a more modest team which loses in the first round of the playoffs? The key to identifying applications like this is to identify a situational variable which one conjectures (or empirically establishes) reliably divides situations into easy-to-lead and hard-to-lead situations. Then, controlling for leadership ability as well as possible, see whether leadership attributions vary in the two kinds of situations.

Of course, there are many obstacles to generalizing our findings. The situations we study are simple analogies to much more complicated organizational choices. While the results of experiment 3 provide support for the persistence of the illusion in more realistic settings, more contextual features can be added. But research often proceeds productively by starting with simple experimental domains, then adding features of context and realism. These experiments are a start, not an end.

In addition, leadership in our experiments is so simple that misattributions are likely to be greater for leaders who display more of the usual characteristics and powers of the situation. What we call "leadership" is a simple, short speech by a randomly chosen group member (more like a spokesperson or temporary leader).[35] As Pfeffer (1977) points out, leaders without many of the usual organizational symbols and powers associated with leadership (such as those in our experiments) are less likely to receive faulty attributions of leadership. Hence, having more realistic forms of leadership is likely to increase the observed effect rather than reduce it.

A final contribution of our research is simply to introduce the fundamental attribution error to research in game theory. In most game theory applications, the game is assumed to be commonly known, so there is little scope for players to make errors in deciding whether outcomes were caused by other players, by chance moves, or by game structure. By using a simple treatment variable, group size, which does affect outcomes to a surprising degree, we create a situational difference for which players can blame others. (In game-theoretic terms, players mistakenly infer from outcomes that the leaders are "low types" or "high types," in terms of leadership skill.) Other experimental games show that attributions of intentions and the fairness of these intentions may matter for outcomes and, hence, for the choices people make.[36] By introducing attribution to game theory and demonstrating its importance, these studies may help bridge the gap between simplified game-theoretic analyses and organizational analyses.

Acknowledgments

This research was funded by NSF grant SBR 95-11001. The authors thank participants at a University of Arizona Department of Management Seminar, the 1997 Public Choice/ESA meetings, the Harvard Behavioral Economics workshop, the 1997 Academy of Management meetings, the Berkeley OB/IR workshop, and the Wharton Conference in honor of Ned Bowman, for helpful comments. The authors are also grateful to Robert Gibbons, Chip Heath, Steve Hoch, R. Mark Isaac, George Loewenstein, Keith Murnighan, Lisa Ordóñez, and several referees for comments and suggestions.

Appendix A. Instructions

Personal Fees and Group Rewards

Set-up & Rules: You have been placed in a group with the other people here today. The group is involved in an activity which has eight rounds. Every round works as follows.

1. Each group member must choose a *personal fee*. The personal fee can be $0.00, $1.00, $2.00, or $3.00.
2. All group members receive the same *reward*. The size of the reward depends on the personal fees you and the other members of the group choose:

If amongst all members of the group the smallest personal fee chosen is	Then you and every other member of the group each receive
a personal fee of $0.00	a reward of $1.00
a personal fee of $1.00	a reward of $2.50
a personal fee of $2.00	a reward of $3.75
a personal fee of $3.00	a reward of $5.00

• Your earnings in a round are simply the reward received minus the personal fee you choose to pay.

Session Leader: One participant will be designated as the session leader. The leader will receive a handout outlining techniques useful in organizing these types of groups. After the second round the leader will address the group and prepare it for the remaining rounds. No other talking is permitted once we begin. Participants violating this rule will be dismissed.

Appendix B. Mathematical Notes

Starting from the following minimization problem:

$$\min_{\hat{\gamma},\hat{\beta}} Var[L^{rated} - L]$$

and recalling that

$$L^{rated} = \hat{\gamma}L^{obs} + \hat{\beta}\left(\frac{P - \alpha_P - \beta_S S}{\beta_L}\right) \quad \text{(A1)}$$

$$L + \tilde{L} = L^{obs} \quad \text{(A2)}$$

$$L + \frac{\tilde{\varepsilon}}{\beta_L} = \frac{P - \alpha_P - \beta_S S}{\beta_L} \quad \text{(A3)}$$

and $Var[X]$ denotes variance of X.

Substituting equation (A1) into the minimization problem and then substituting in the left hand term in Equations (A2) and (A3) gives:

$$\min_{\hat{\gamma},\hat{\beta}} Var\left[\hat{\gamma}(L + \tilde{L}) + \hat{\beta}\left(L + \frac{\tilde{\varepsilon}}{\beta_L}\right) - L\right]$$

Recalling that $Cov[L, \tilde{L}] = Cov[L, \tilde{\varepsilon}] = Cov[\tilde{L}, \tilde{\varepsilon}] = 0$ this is equal to

$$\min_{\hat{\gamma},\hat{\beta}} (\hat{\gamma} + \hat{\beta} - 1)^2 Var[L] + \hat{\gamma}^2\sigma_{\tilde{L}}^2 + \frac{\hat{\beta}^2}{\beta_L^2} Var[\tilde{\varepsilon}]$$

Taking the first-order conditions with respect to $\hat{\gamma}$ and $\hat{\beta}$ we get

$$2(\hat{\gamma} + \hat{\beta} - 1)\, Var[L] + \frac{2\hat{\beta}}{\beta_L^2}\sigma_{\tilde{\varepsilon}}^2 = 0 \quad \text{(A4)}$$

$$2(\hat{\gamma} + \hat{\beta} - 1)\, Var[L] + 2\hat{\gamma}\sigma_{\tilde{L}}^2 = 0 \quad \text{(A5)}$$

Solving for $\hat{\gamma}$ and $\hat{\beta}$ gives:

$$\hat{\beta} = \frac{1}{1 + \frac{\sigma_{\tilde{\varepsilon}}^2}{\beta^2 L \sigma_{\tilde{L}}^2} + \frac{\sigma_{\tilde{\varepsilon}}^2}{\beta_L^2\, Var[L]}} \quad \text{(A6)}$$

$$\hat{\gamma} = \frac{1}{1 + \frac{\beta_L^2\sigma_{\tilde{L}}^2}{\sigma_{\tilde{\varepsilon}}^2} + \frac{\sigma_{\tilde{L}}^2}{Var[L]}} \quad \text{(A7)}$$

Endnotes

[1]See, for instance, Meindl et al. 1985, Meindl and Ehrlich 1987, Mitchell et al. 1977, Lord et al. 1978, and Staw and Ross 1980. Most of these studies use what we call "the performance cues paradigm," in which subjects are exposed to a part of a leadership situation, the performance of the group being led is exogenously manipulated, and subjects are asked to rate the leaders. We discuss this research and some problems we believe exist with this method extensively in a later section.

[2]The universality of the tendency to overweight personal variables compared to situational variables has been questioned. For example, Morris and Peng (1994) provide evidence that Chinese high school students, graduate students, and reporters weight situational variables more heavily than do Americans. Furthermore, there is evidence that telling subjects that they will subsequently have to justify and explain their actions reduces the attribution effect (Tetlock 1985).

[3]This has also been studied experimentally by Cachon and Camerer (1996), Knez and Camerer (1994, 1996); Camerer et al. (1996); and theoretically by Anderson et al. (1996) and Crawford (1995).

[4]See Camerer and Knez (1997) and Camerer (in progress) for experimental evidence.

[5]Additional examples of weak-link games include: a group trying to assemble a project report on time (like contributing chapters to an edited book); people meeting in a restaurant who cannot sit until the last person arrives (assuming they prefer to sit than to wait around), and "high-reliability" organizations that demand high levels of safety input by each member.

[6]There are no significant differences between the two subject populations, and hence, the data are pooled.

[7]Instructions are in the appendix.

[8]Subjects also got a handout of the Table 1 matrix. They were told that this matrix summarized the payoffs described in the instructions. However, several cells were left blank and subjects were instructed to fill in the correct numbers. The experimenter checked to make sure that all subjects filled in the cells correctly, and then publicly went through each of the calculations. This was done to ensure that subjects knew how to calculate the earnings associated with each outcome and to make this common knowledge.

[9]VHBB (1990) showed that giving full information about the distribution of personal fees at the end of each round, instead of just the group minimum, did not make much difference.

[10]We gave them a standard speech to control for differences in leadership ability.

[11]In addition, if the leaders were allowed to say and do more, then it is likely that true differences in leadership ability would emerge so that we are not holding leadership constant. If one had a large sample of leaders we could safely assume that true leadership ability is about the same across large and small groups, on average, but large samples are expensive because only one leader is used in each large group. An obvious alternative is to use deception and employ confederate leaders who have a wider range of leaderlike actions available to them, but make exactly the same speech. Like most experimental economists, we prefer to avoid deception unless absolute necessary (see Camerer 1996 for further discussion).

[12]We did not conduct control groups without leaders to see how group fees would have changed over time. The results from several studies mentioned earlier strongly suggest that small groups will converge to the efficient outcome, and large groups will converge to the inefficient outcome, but we do not know if this is necessarily true in this setting. However, a leaderless control group is not necessary for interpreting our results because we are interested in the difference between large and small groups rather than the difference between groups with and without leaders (holding group size constant). The latter comparison

would be useful for judging whether groups with a leader blame other group members more or less when a leader is present—that is, does the leader act as a "blame sponge" or "blame lightning rod," which reduces blame attributed to other group members? However, this question extends beyond the scope of our paper.

[13]We omit the leaders' responses, which reduces n to 35 and 36 for large groups and pairs, respectively.

[14]This significance is at the $p < 0.005$ level by a t-test ($t_{67} = 3.352$). Furthermore, the difference between the distribution of ratings between the two conditions is significant at the $p < 0.01$ by an Epps-Singleton (1986) characteristic function test ($\chi^2_4 = 14.255$).

[15]Baron and Hershey (1988) discuss a related phenomenon known as outcome bias: When subjects are asked to rate the quality of thinking behind a decision made by others under conditions of uncertainty, they rate the thinking as better when the outcome is favorable than when it is unfavorable. To the extent that players in the pairs condition experience favorable outcomes while players in the large groups do not, outcome bias predicts a difference in retrospective evaluation of the leaders. Note, however, that in our set-up the difference in outcomes is not due to the resolution of uncertainty, but to either the personality of the leaders or the nature of the situation in which they are placed. We ask subjects to tell us which of the two is actually responsible for their outcomes, and they answer incorrectly.

[16]That is, the fundamental attribution error is not truly fundamental, but is a by-product of the fact that actions of people are "fundamentally" more salient and mutable. This implies that longer and more elaborate speeches by the leaders are likely to lead to a larger attribution effect.

[17]Notice that by charging those who vote to "fire" the leader a small $0.25 fee, our test for attribution error is in fact quite conservative: Even if subjects think the leader should be replaced, if they think that a majority of other subjects will vote to replace the leader they can vote to keep him and save the fee. Thus, voting probably understates the strength of preference to get rid of a bad leader, biasing the results against our hypothesis.

[18]Support for this conjecture can be found in Tables 4 and 6, which show that early personal fee choices in Experiment 2 differ across large and small groups more than they do in Experiment 1, indicating that subjects better anticipate the group-size effect from the very start. This is evidence that the misattributions to leadership quality persist even when subjects are more *strongly* aware of the group-size effect than they were in Experiment 1.

[19]In addition, a comparison of the mean rating given to the leader by those who voted to replace (3.06) with the mean rating given by those who voted to keep (4.58) reveals a difference significant at the $p < 0.001$ level using a one-tailed t-test ($t_{73} = 3.638$).

[20]The pairs condition consisted of four pairs (eight subjects) while the large groups were comprised of eight and nine subjects.

[21]The choices in this table represent the week during which a subject chose to submit his or her part of the report. A choice of 4 means that the subject waited until the last week and corresponds to the inefficient equilibrium (personal fee = $0), while a choice of 1 means that the subject completed his or her part in the first week and corresponds to the efficient equilibrium (personal fee = $3). The personal fee corresponding to each choice is given in parentheses.

[22]Meindl and his coauthors are careful to point out, however, that even if leadership is romanticized and perceptions are attribution errors, these perceptions are worth studying in their own right (see also Calder (1977); Pfeffer (1977)). Organizational researchers should not abandon the study of leadership entirely, but instead shift toward a "follower-centric" view.

[23]There is also a restriction of range problem in these studies: If only the best leaders become CEOs or mayors, then the estimated correlation between leadership ability and performance is lower than it would be in a population of randomly chosen leaders.

[24]Many studies rate behavior using the Leader Behavior Description Questionnaire (LBDQ) (see Stogdill 1963).

[25]Chen and Meindl (1991) used content analysis of business press writings to show that the image of a corporate leader in the popular press rose and fell according to the fortunes of his company. Also using nonexperimental data, Meindl et al. (1985) reported a positive linear relation between company performance and the relative frequency of leadership mentions in press coverage of those companies. (However, note that there is no linear relation in their experimental findings, which are U-shaped.) They also reported that the frequency of dissertations written on leadership fell with lagged increases in GNP while, oppositely, frequency of leadership articles in the business press rose with lagged increases in GNP.

[26]This argument, as well as the following mathematical example, is similar to the one underlying Dawes' (1990) claim that the false consensus effect (the tendency to correlate one's own response with predictions of responses of others) is not false if subjects are treating their own response as a diagnostic observation of the population's propensity, as Bayesian principles would suggest.

[27]Assume that $\tilde{L}$ is distributed $N[0, \sigma^2_{\tilde{L}}]$ and that $Cov[L, \tilde{L}] = 0$.

[28]Assume that $\tilde{\varepsilon}$ is distributed $N[0, \sigma^2_{\tilde{\varepsilon}}]$ and that $Cov[L, \tilde{\varepsilon}] = Cov[\tilde{L}, \tilde{\varepsilon}] = 0$.

[29]In Staw (1975) subjects did not rate leadership skill, but the effect of the performance cue on ratings of group behaviors and qualities was vastly more significant when outside subjects rated based on a vignette, than when subjects who participated in the task rated their own groups.

[30]As an anonymous referee pointed out, this is again similar to Dawes' argument concerning the false consensus effect in that subjects are getting the direction right in their responses and previous research has not shown that the magnitude is incorrect.

[31]Many social psychologists are critical of pencil-and-paper studies or those which don't immerse subjects in a live situation, often because it is difficult to tell what subjects should be thinking and whether the attributions are incorrect (e.g., Jones and Nisbett, 1971). As one prominent social psychologist put it, "You can't do social psychology without creating social reality."

[32]One possible explanation of the results in Experiment 1 is that outcome bias, the tendency to rate previous thinking as better when it leads to better outcomes than when it does not, is producing the difference between the leaders' rating once subjects have observed all of the outcomes. There are two reasons, however, why outcome bias cannot be the only explanation of the results. First, we conducted an additional set of experiments (not reported in this paper) which deal with succession in leadership. We found that when successful leaders were followed by leaders who were equally (or even more) successful, the successors were not rated as highly as their predecessors. Since the second leaders in these experiments produce results which are at least as good as those of the first leaders, outcomes are "held fixed" and outcome bias alone would predict that the second leaders would receive

equal ratings. In fact, final ratings are lower, supporting the idea that attribution, and not outcome bias, produces the result in Experiment 1. Second, subjects in Experiment 1 blamed other subjects as well as the leader for failure in the large groups. If outcome bias were the only explanation of our result, we might expect that the blame placed on other subjects would be equal to the blame placed on the leaders. This is not true: the leaders are held more accountable for the group's success or failure, indicating that attributions of leadership quality are being made. Hence, while we do not claim that outcome bias has nothing to do with the results, it is not the only cause of the difference in ratings.

[33]The leaders' questionnaire responses have been omitted from the analysis thus far. An inspection of their self-ratings, however, reveals that they are prone to the same attribution errors. The mean initial ratings (pooling across experiments) are 6.2 for pairs and 6.0 for large groups, and this difference is not significant at any reasonable levels. The final ratings, however, are 7.4 for pairs and 4.4 for large groups and this difference is significant at the $p < 0.005$ level ($t_8 = 4.161$). There are only a few observations (five in each condition), however, because several leaders did not completely fill out the questionnaires.

[34]Readers will appreciate how difficult it can be to speak across intellectual boundaries—say, between sociology, psychology, and economics—even when the speakers all use a common language, say English. One problem is that a common term, like "rational" or "theory," can be used very differently in different disciplines. The opposite problem is that fields will develop specialized terms, like "operant conditioning," "two-stage least-squares," or "munificence," which mean nothing to English-speakers in other disciplines. Both problems limit communication.

[35]Obviously, more elaborate sorts of leadership could be created in the experiments. A danger with doing so, however, is that true differences in leadership could emerge and would affect results, so that establishing a misattribution of leadership ability is difficult. By keeping the nature of the leadership activity short and simple, we ensure (as the postspeech ratings show) that there are few differences in perceived leadership in the small and large group, so situational differences loom large and can be misattributed to people.

[36]Blount (1995) reports that in ultimatum bargaining, players are willing to accept lower offers from a random device than from another player who benefits from the lower offer. Gibbons and Van Boven (1997) show that subjects make attribution errors about the other player in the prisoner's dilemma on the basis of an essay for or against cooperation, even when told that the other player had no choice concerning which side to take. Schotter et al. (1995) show that competitive pressures placed on proposers in ultimatum games lead to offers which are lower and slightly more likely to be accepted, indicating that responders' considerations of fairness may be affected by whether they believe proposers are making low offers because they are selfish or because they have to. Rabin (1993) constructs a game-theoretic model of "fairness equilibrium" in which players form beliefs about another player's "intentions," then reciprocate perceived niceness of players who they think intend to be nice, and similarly for perceived meanness (see also Geanakopolos et al. 1989). In Rabin's model, cooperation in the prisoners' dilemma may be reciprocated (it is a "nice" action), but if the other person is forced to cooperate (e.g., by law or an outside force), then that person's cooperation is not nice and will not be reciprocated

References

Anderson, S. P., J. K. Goeree, C. A. Holt. 1996. Minimum-effort coordination games: An equilibrium analysis of bounded rationality. Unpublished manuscript.

Baron, J., J. C. Hershey. 1988. Outcome bias in decision evaluation. *J. Personality and Social Psych.* **54**(4) 569–579.

Blount, S. 1995. When social outcomes aren't fair: The effect of causal attributions on preferences. *Organ. Behavior and Human Decision Process.* **63**(2) 131–144.

Brockner, J., B. M. Wiesenfeld, C. L. Martin. 1995. Decision frame, procedural justice and survivors' reactions to job layoff. *Organ. Behavior and Human Decision Process.* **63**(1) 59–68.

Cachon, G. P., C. F. Camerer. 1996. Loss avoidance and forward induction in experimental coordination games. *Quart J. Econom.* **111** 165–194.

Calder, B. J. 1977. An attribution theory of leadership. B. M. Staw, G. R. Salancik, eds. *New Directions in Organizational Behavior* St. Clair Press, Chicago, IL.

Camerer, C. F. 1996. How to do experiments in economics and psychology, and why they differ. D. Budescu, I. Erev, R. Zwick, eds. *Experiments on Strategic Interaction: Essays in Honor of Reinhard Selten* Springer-Verlag, Berlin, Germany.

——. Experiments on behavioral game theory. Book manuscript in progress.

——, M. Knez, R. A. Weber. 1996. Timing and virtual observability in ultimatum bargaining and weak-link coordination games. Working paper No. 970, California Institute of Technology, Pasadena, California.

——, ——. 1997. Coordination in organizations: A game-theoretic perspective. Zur Shapira, ed. *Organizational Decision Making* Cambridge Series on Judgment and Decision Making, Cambridge, U.K.

Chen, C. C., J. R. Meindl. 1991. The construction of leadership images in the popular press: The case of Donald Burr and People Express. *Admin. Sci. Quart.* **36** 521–551.

Crawford, V. P. 1995. Adaptive dynamics in coordination games. *Econometrica* **63**(1) 103–143.

Dawes, R. M. 1990. The potential nonfalsity of the false consensus effect. R. M. Hogarth, ed. *Insights in Decision Making: A Tribute to Hillel J. Einhorn* The University of Chicago Press, Chicago, IL..

Epps, T. W., K. J. Singleton. 1986. An omnibus test for the two-sample problem using the empirical characteristic function. *J. Statist. Computational Simulations* **26** 177–203.

Geanakopolos, J., D. Pearce, E. Stacchetti. 1989. Psychological games and sequential rationality. *Games and Econom. Behavior* **1** 60–79.

Gibbons, R., L. Van Boven. 2001. Contingent social utility in the "Prisoner's Delemma." *J. Econom. Behavior Organ.* **45**(1) 1–17.

Jones, E. E., R. E. Nisbett. 1971. The actor and the observer: Divergent perceptions of the causes of behavior. E. E. Jones, D. E. Kanouse, H. H. Kelley, R. E. Nisbett, S. Valins, and B. Weiner, eds. *Attribution: Perceiving the Causes of Behavior* General Learning Press, New York.

Kahneman, D., D. T. Miller. 1986. Norm theory: Comparing reality to its alternatives. *Psych. Rev.* **93** 136–153.

Knez, M., C. F. Camerer. 1994. Creating expectational assets in the laboratory: Coordination in weakest-link games. *Strategic Management J.* **15** 101–119.

——, ——. 1996. Increasing cooperation in social dilemmas through the precedent of efficiency in coordination games. Working paper, University of Chicago, Chicago, IL.

——, D. Simester. 2001. Firmwide incentives and mutual monitoring at Continental Airlines. *J. Labor Econom.* Forthcoming.

Kogut, B., N. Kulatilaka. 2001. Capabilities as real options. *Organ. Sci.* Forthcoming.

Larrick, R. P., M. W. Morris. 1995. When one casts doubt on another: A normative analysis of discounting in causal attribution. *Psych. Rev.* **102**(2) 331–355.

Lieberson, S., J. F. O'Connor. 1972. Leadership and organizational performance: A study of large corporations. *Amer. Soc. Rev.* **37** 117–130.

Lord, R. G., J. F. Binning, M. C. Rush, J. C. Thomas. 1978. The effect of performance cues and leader behavior on questionnaire ratings of leadership behavior. *Organ. Behavior and Human Performance* **21** 27–39.

Meindl, J.R. 1995. The romance of leadership as a follower-centric theory: A social constructionist approach. *Leadership Quart.* **6**(3) 329–341.

——, S. B. Ehrlich. 1987. The romance of leadership and the evaluation of organizational performance. *Acad. Management J.* **30**(1) 91–109.

——, S. E. Ehrlich, J. M. Dukerich. 1985. The romance of leadership. *Admin. Sci. Quart.* **30** 78–102.

Milgrom, P. R., J. Roberts. 1992. *Economics, Organization, and Management.* Prentice-Hall, Englewood Cliffs, NJ.

Mitchell, T. R., J. R. Larson, Jr., S. G. Green. 1977. Leader behavior, situational moderators, and groups performance: An attributional analysis. *Organ. Behavior and Human Performance* **18** 254–268.

Morris, M. W., K. Peng. 1994. Culture and cause: American and Chinese attributions for social and physical events. *J. Personality and Soc. Psych.* **67** 949–971.

Murnighan, J. K., J. W. Kim, A. R. Metzger. 1993. The volunteer dilemma. *Admin. Sci. Quart.* **38**(4) 515–538.

Pfeffer, J. 1977. The ambiguity of leadership. *Acad. Management Rev.* **2** 104–112.

Accepted by Arie Lewin.

Rabin, M. 1993. Incorporating fairness into game theory and economics. *Amer. Econom. Rev.* **83**(5) 1281–1302.

Ross, L., T. M. Amabile, J. L. Steinmetz. 1977. Social roles, social control, and biases in social perception. *J. Personality and Soc. Psych.* **32** 880–892.

——, R. E. Nisbett. 1991. The Person and the Situation. McGraw-Hill, New York.

Salancik, G. R., J. Pfeffer. 1977. Constraints on administrative discretion: The limited influence of mayors on city budgets. *Urban Affairs Quart.* **12** 475–498.

Schotter, A., A. Weiss, I. Zapater. 1996. Fairness and survival in ultimatum and dictatorship games. *J. Econom. Behavior and Organ.* **31** 37–56.

Simon, H. A. 1991. Organizations and markets. *J. Econom. Perspectives* **5**(2) 25–44.

Staw, B. M. 1975. Attribution of the 'causes' of performance: A general alternative interpretation of cross-sectional research on organizations. *Organ. Behavior and Human Performance* **13** 414–432.

——, J. Ross. 1980. Commitment in an experimenting society: A study of the attribution of leadership from administrative scenarios. *J. Appl. Psych.* **65**(3) 249–260.

Stogdill, R. M. 1948. Personal factors associated with leadership: A survey of the literature. *J. Psych.* **25** 35–71.

——. 1963. *Manual for the Leader Behavior Description Questionnaire.* Bureau of Business Research, Ohio State University, Columbus, OH.

Taylor, S. E., S. T. Fiske. 1975. Point-of-view and perceptions of causality. *J. Personality and Soc. Psych.* **32** 439–445.

Tetlock, P. E. 1985. Accountability: A social check on the fundamental attribution error. *Soc. Psych. Quart.* **48**(3) 227–236.

Thompson, J. D. 1967. Organizations in Action. McGraw-Hill, New York.

Tushman, M. L., J. P. Murmann. 1998. Dominant designs, innovation types, and organizational outcomes. *Res. Organ. Behavior* **20** 231–266.

Van Huyck, J. B., R. C. Battalio, R. O. Beil. 1990. Tacit coordination games, strategic uncertainty and coordination failure. *Amer. Econom. Rev.* **80**(1) 234–248.

Winter, S. G., G. Szulanski. 2001. Replication as strategy. *Organ. Sci.* Forthcoming.