Leadership in Groups: A Monetary Policy Experiment*

Alan S. Blinder[a] and John Morgan[b]
[a]Princeton University
[b]University of California, Berkeley

This paper studies monetary policy decision making by committee, using an experimental methodology. In an earlier paper (Blinder and Morgan 2005), we found that groups not only outperformed individuals, but they also took no longer to reach decisions. We successfully replicate those results here. Next, we find little difference between the performances of four-person and eight-person groups; the larger groups outperform the smaller groups by a very small (and often insignificant) margin. Third, and most surprisingly, we find no evidence of superior performance by groups that have designated leaders. Possible reasons for that strongly counterintuitive finding are discussed.

JEL Codes: C92, E58.

1. Introduction and Motivation

The transformation of monetary policy decisions in most countries from individual decisions to group decisions is one of the most notable developments in the recent evolution of central banking (Blinder 2004, ch. 2). In an earlier paper (Blinder and Morgan 2005), we ran an experiment in which Princeton University students, acting as ersatz central bankers, made monetary policy decisions both as individuals and in groups. Those experiments yielded two main findings:

*We are grateful to Jennifer Brown, Jae Seo, and Patrick Xiu for fine research assistance and to the National Science Foundation and Princeton's Center for Economic Policy Studies for financial support. We also acknowledge extremely helpful comments from Petra Geraats, Petra Gerlach-Kristen, Jens Grosser, Helmut Wagner, a referee, and seminar participants at Princeton, the International Monetary Fund, and the National Bureau of Economic Research.

(i) Groups made better decisions than individuals, in a sense to be made precise below.
(ii) Groups took no longer to reach decisions than individuals did.[1]

The first finding was not a big surprise, given the previous literature on group versus individual decision making (most of it from disciplines other than economics). But we were frankly stunned by the second finding. Like seemingly everyone, we believed that groups moved more slowly than individuals. A subsequent replication with students at the London School of Economics (Lombardelli, Proudman, and Talbot 2005) verified the first finding but did not report on the second one.

This paper replicates our 2005 findings using the same experimental apparatus, but with students at the University of California, Berkeley. That the replication is successful bolsters our confidence in the Princeton results. But that is not the focus of this paper. Instead, we study two important issues that were deliberately omitted from our previous experimental design.

The first pertains to *group size*. In the Princeton experiment, every monetary policy committee (MPC) had five members—precisely (and coincidentally) the size that Sibert (2006) subsequently judged to be optimal. Lombardelli, Proudman, and Talbot (2005), following our lead, also used committees of five. But real-world monetary policy committees vary greatly in size, so it seems important to compare the performance of small versus large groups. Revealed-preference arguments offer little guidance in this matter, since real-world MPCs range in size from three to twenty-one, with the European Central Bank (ECB) headed even higher. In this paper, we study the size issue by comparing the experimental performances of groups of four and eight.[2]

[1]In both our 2005 paper and the present one, "time" is measured by the amount of *data* required before the individual or group decides to change the interest rate—not by the number of ticks of the clock. Our reason was (and remains) simple: this is the element of time lag that is relevant to monetary policy decisions; no one cares about how many hours the committee meetings last.

[2]The reason for choosing even-numbered groups will be made clear shortly. Our "large" groups ($n = 8$) are still small compared with, e.g., the ECB or the Federal Reserve. This size was more or less dictated by the need to recruit large

The second issue pertains to *leadership* and is the unique aspect of the research reported here. Both our Princeton experiment and Lombardelli, Proudman, and Talbot's replication treated all members of the committee equally. But every real-world monetary policy committee has a designated leader who clearly outranks the others. At the Federal Reserve, that leader is the "chairman"; at the ECB, he is the "president"; and at the Bank of England and many other central banks, he or she is the "governor." Indeed, we are hard-pressed to think of *any* committee, in *any* context, that does *not* have a well-defined leader. Juries come close, but even they have foremen. Observed reality, therefore, strongly suggests that groups need leaders in order to perform well. But is it true? That is the main question that motivates this research.

Consider leadership on MPCs in particular. While all MPCs have designated leaders, the leader's authority varies greatly. The Federal Open Market Committee (FOMC) under Alan Greenspan (but not under Ben Bernanke) was at one extreme; it was what Blinder (2004, ch. 2) called an *autocratically collegial* committee, meaning that the chairman came close to dictating the committee's decision. This tradition of strong leadership did not originate with Greenspan. Paul Volcker's dominance was legendary, and Chappell, McGregor, and Vermilyea (2005, ch. 7) estimated econometrically that Arthur Burns's views on monetary policy carried about as much weight as those of all other FOMC members combined. At the other extreme, the Bank of England's MPC is what Blinder (2004) called an *individualistic* committee—one that reaches decisions (more or less) by true majority vote. Governor Mervyn King has even allowed himself to be outvoted, partly in order to make this point. In between these poles, we find a wide variety of *genuinely collegial* committees, like the ECB Governing Council, that strive for consensus. Some of these committees are led firmly; others are led only gently.

The scholarly literature on group decision making, which comes mostly from psychology and organizational behavior, offers relatively little guidance on what to expect. And only a small portion of it is experimental. As a broad generalization, our quick review of the literature led us to expect to find some positive effects of leadership on

numbers of subjects. With groups of four and eight, we needed 252 subjects in all.

group performance—which is the same prior we had before reviewing the literature. But it also led to some doubts about whether intellectual ability is a key ingredient in effective leadership (Fiedler and Gibson 2001). Instead, the literature suggests that gains from group interaction may depend more on how well the leader encourages other members of the group to contribute their opinions frankly and openly (Maier 1970, Blades 1973, and Edmondson 1999). In an interesting public goods experiment, Guth et al. (2004) found that stronger leadership produced better results, although the leaders in that experiment were selected randomly. We did not find any relevant evidence on whether leadership effects are greater in larger or smaller groups.

With these two issues—group size and leadership—in mind, we designed our experiment with four treatments, running ten or eleven sessions with each treatment:

(i) four-person groups with no leader, hereafter denoted $\{n = 4,$ no leader$\}$
(ii) four-person groups with a leader $\{n = 4,$ leader$\}$
(iii) eight-person groups with no leader $\{n = 8,$ no leader$\}$
(iv) eight-person groups with a leader $\{n = 8,$ leader$\}$

We summarize our results briefly here because they will be understood far better after the experimental details are explained. First, we successfully replicate our Princeton results, at least qualitatively: groups perform better than individuals, and they do not require more "time" to do so. Second, we find little difference between the performance of four-person and eight-person groups; the larger groups outperform the smaller groups by a very small (and often insignificant) margin. Third, and most important, we find no evidence of superior performance by groups that have designated leaders. Groups without such leaders do as well as or better than groups with well-defined leaders. This is a surprising finding, and we will speculate on some possible reasons later.

The rest of the paper is organized as follows. Section 2 describes the experimental setup, which in most respects is exactly the same as in Blinder and Morgan (2005). Section 3 briefly presents results comparing group and individual performance that mostly replicate those of our Princeton experiment. Sections 4 and 5 focus on the

data generated by decision making in groups, presenting new results on the effects of group size and leadership, respectively. Then section 6 summarizes the conclusions.

2. The Experimental Setup[3]

Our experimental subjects were Berkeley undergraduates who had taken at least one course in macroeconomics. We brought them into the Berkeley Experimental Social Sciences Lab (Xlab) in groups of either four or eight, telling them only that they would be playing a monetary policy game. Except by coincidence, the students did not know one another beforehand. Each computer was programmed with the following simple two-equation macroeconomic model—exactly the same one used in the Princeton experiment—with parameters chosen to resemble the U.S. economy:

$$\pi_t = 0.4\pi_{t-1} + 0.3\pi_{t-2} + 0.2\pi_{t-3} + 0.1\pi_{t-4} - 0.5(U_{t-1} - 5) + w_t \tag{1}$$

$$U_t - 5 = 0.6(U_{t-1} - 5) + 0.3(i_{t-1} - \pi_{t-1} - 5) - G_t + e_t. \tag{2}$$

Equation (1) is a standard accelerationist Phillips curve. Inflation, π, depends on the deviation of the lagged unemployment rate from its presumed natural rate of 5 percent, and on its own four lagged values, with weights summing to one. The coefficient on the unemployment rate is chosen roughly to match empirically estimated Phillips curves for the United States.

Equation (2) can be thought of as an IS curve with the unemployment rate, U, replacing real output (via Okun's Law). Unemployment tends to rise above (or fall below) its natural rate when the *real* interest rate, $i - \pi$, is above (or below) its "neutral" value, which is also set at 5 percent. (Here i is the nominal interest rate.) But there is a lag in the relationship, so unemployment responds to the real interest rate only gradually. Like real-world central bankers, our experimental subjects control only the *nominal* interest rate, not the *real* interest rate.

[3]This section overlaps substantially with section 1.1 of Blinder and Morgan (2005) but omits some of the detail presented there.

The G_t term in (2) is the shock to which our student monetary policymakers are supposed to react. It starts at zero and randomly changes *permanently* to either $+0.3$ or -0.3 sometime during the first ten periods of play. Readers can think of G as representing government spending or any other shock to aggregate demand. As is clear from (2), a change in G changes U by precisely the same amount, but in the opposite direction, on impact. Then there are lagged responses, and the model economy eventually converges back to its natural rate of unemployment. Because of the vertical long-run Phillips curve, any constant inflation rate paired with $U = 5\%$ can be an equilibrium.

We begin each round of play with inflation at 2 percent—which is also the central bank's target rate (see below). Thus, prior to the shock (i.e., when $G = 0$), the model's steady-state equilibrium is $U = 5$, $i = 7$, $\pi = 2$. As is apparent from the coefficients in equation (2), the shock changes the neutral real interest rate from 5 percent to either 6 percent or 4 percent *permanently*. Our subjects—who do *not* know this—are supposed to detect and react to this change, presumably with a lag, by raising or lowering the nominal interest rate accordingly.

Finally, the two stochastic shocks, e_t and w_t, are drawn independently from uniform distributions on the interval $[-.25, +.25]$.[4] Their standard deviations are roughly half the size of the G shock. This sizing decision, we found, makes the fiscal shock relatively easy to detect—but not too easy.

Lest our subjects had forgotten their basic macroeconomics, the instructions remind them that raising the interest rate lowers inflation and raises unemployment, while lowering it does the reverse, albeit with a lag.[5] In the model, monetary policy affects unemployment with a one-period lag and inflation with a two-period lag; but students are not told that. Nor are they told anything else about the model's specification. They *are* told that the demand shock will occur at a random time that is equally likely to be any of periods 1 through 10. But they are told neither the magnitude of this shock, nor its direction, nor whether it is permanent or temporary.

[4]The distributions are uniform, rather than normal, for programming convenience.

[5]The instructions are provided in the appendix.

Doubtless, this little model economy is far simpler than the actual economies that real-world central bankers try to manage. However, *to the student subjects*, who do not know anything about the model, we believe this setup poses perplexities that are comparable to those facing real-world central bankers, who are trying to stabilize a much more complex system (e.g., one that includes expectational effects) but who also know much more, have far more experience, and have abundant staff support. For example, our experimental subjects do not know the transmission mechanism, the lag structure, whether the price equation is forward looking or backward looking, and so on. Nor do they have the benefit of staff forecasts or analyses.

Despite the model's seeming simplicity, stabilizing it can be tricky in practice. Because of the unit root apparent in equation (1), the model diverges from equilibrium when perturbed by a shock—unless it is stabilized by monetary policy. But lags and modest early-period effects combine to make the divergence from equilibrium pretty gradual and hence less than obvious at first. Once unemployment and inflation start to "run away from you," it can be difficult to get them back on track. Furthermore, it is not easy to distinguish between the permanent G shock and the transitory e and w shocks that add "noise" to the system. Indeed, the subjects do not even know that the G shock is permanent while the others are i.i.d.

Each play of the game proceeds as follows. We start the system in steady-state equilibrium at the values mentioned above. The computer then selects values for the two random shocks and displays the first-period values of U and π, which are typically quite close to the target values ($U = 5\%$, $\pi = 2\%$), on the screen for the subjects to see. In each subsequent period, new random values of e_t and w_t are drawn, thereby creating statistical noise, and the lagged variables that appear in equations (1) and (2) are updated. At some random time, unknown to subjects, the G shock occurs. The computer calculates U_t and π_t each period and displays them on the screen, where all past values are also shown. Subjects are then asked to choose an interest rate for the next period, and the game continues for twenty such periods. Students are told to think of each period as a quarter, so the simulation covers "five years." Each five-year run of the game is different because of different random draws.

No time pressure is applied; subjects are permitted to take as much clock time as they wish to make each decision. As noted above, the concept of time that interests us is the *decision lag*: the amount of *new data* the decision maker insists upon before changing the interest rate. In the real world, data flow in unevenly over calendar time; in our experiment, subjects see exactly one new observation on unemployment and inflation each period. So when we say later that one type of decision-making process "takes longer" than another, we mean that more *data* (not more *minutes*) are required.

To rate the quality of their performances, and to reward subjects accordingly, we tell students that their score for each quarter is

$$s_t = 100 - 10|U_t - 5| - 10|\pi_t - 2|, \tag{3}$$

and the score for the entire game (henceforth, S) is the (unweighted) average of s_t over the twenty quarters. We use an absolute-value function instead of the quadratic loss function that is ubiquitous in research on monetary policy (and elsewhere) because quadratics are too hard for subjects—even Princeton and Berkeley students—to calculate in their heads. Notice also that the coefficients in equation (3) scale the scores into percentages, which gives them a natural, intuitive interpretation. Thus, e.g., missing the unemployment target by 0.8 (in either direction) and the inflation target by 1.0 results in a score of $100 - 8 - 10 = 82$ (percent) for that period.[6] At the end of the session, scores are converted into money at the rate of 25¢ per percentage point. Subjects typically scored 80–84 percent of the possible points, thus earning about \$20–\$21.

One final detail needs to be mentioned. To deter excessive manipulation of the interest rate (which we observed in testing the apparatus in dry runs), we charge subjects a fixed cost of 10 points each time they change the rate of interest, regardless of the size of the change.[7] Ten points is a small charge; averaged over a twenty-period game, it amounts to just 0.5 percent of the total potential score. But we found it to be large enough to deter most of the excessive fiddling with interest rates. Analogously, researchers who try to derive

[6]The unemployment and inflation data are always rounded to the nearest tenth. So students see, e.g., 5.8 percent, not, say, 5.83 percent.

[7]To keep things simple, only integer interest rates are allowed.

the Federal Reserve's reaction function from the minimization of a quadratic loss function find that they must add something like $k(i_t - i_{t-1})^2$ to the loss function in order to fit the data. Without that term, interest rates turn out to be far more volatile than they are in practice.[8]

The sessions are played as follows. Either four or eight students enter the lab and are read detailed instructions, which they are also given in writing. (See the appendix.) In the case of sessions with a designated leader, the instructions tell them, among other things, that the person earning the highest score while playing alone in part 1 of the experiment will be designated the "leader" (the term we use) of the group for part 2—where he or she will be rewarded with a doubled score.

Subjects are then allowed to practice with the computer apparatus for five minutes, during which time they can ask any questions they wish. At the end of the practice period, each machine is reinitialized, and each student is instructed to play twelve rounds of the game (each lasting twenty "quarters") *alone*—without communicating in any way with the other subjects. Once all the subjects have completed twelve rounds of individual play, the experimenter calls a halt to part 1 of the experiment.

In part 2, the same students gather around a single large screen to play the same game twelve times *as a group*. It is here that the sessions with and without leaders differ. In leaderless sessions, the rules are exactly the same as in individual play, except that students are now permitted to communicate freely with one another—as much as and in any way they please. Everyone in the group is treated alike, and each subject receives the group's common score.

In sessions with a designated leader, the experimenter begins by revealing who earned the highest score in part 1, and that student becomes the leader for part 2.[9] Thus, the criterion for electing leaders is purely intellective: the skill of an individual at ersatz monetary policymaking. Since the group will perform the identical task, this selection principle would seem a natural one.

[8] See, e.g., Rudebusch (2001).

[9] On average, that student scored 10.8 points higher than the others in the group during part 1 of the experiment, a sizable difference. But students were not told the leader's score.

Table 1. The Flow of the Experiment

Instructions
Practice Rounds (no scores recorded)
Part 1: Twelve rounds played as individuals
Part 2: Twelve rounds played as a group (with or without a leader)
Part 3: Twelve rounds played as individuals
Students are paid by check and leave

The meaning of leadership in the experiment is threefold: First, the leader is responsible for communicating (verbally) the group's decision to the experimenter—which makes the leader pivotal to the discussion. Second, the leader faces stronger incentives: his or her score in part 2 is *double* that of the other subjects. Third, the leader gets to break any tie vote—which is why we use even-numbered groups.[10] While we recognize that the experimental setup still allows limited scope for leadership, we judged that this was about all we could do in a laboratory setting with $1^1/_2$ hours of experimental time. We return to this issue later.

After twelve rounds of group play, the subjects return to their individual computers for part 3, in which they play the game another twelve times alone, with no communication with the others. For future reference, table 1 summarizes the flow of each session.

A typical session (of 36 rounds of the game) lasted about 90 minutes, and we ran 42 sessions in all, amounting to 252 total subjects. (No subject was permitted to play more than once.) Each of the 21 four-person sessions *should have* generated 24 individual rounds of play per subject, or $21 \times 4 \times 24 = 2,016$ in all, plus 12 group rounds per session, or 252 in all. Each of the 21 eight-person sessions *should have* generated twice as many individual observations (hence 4,032 in total), plus another 252 group observations. Thus we have a plethora of data on individual performance but a relative paucity of data on group performance. Since a small number of observations were lost due to computer glitches, table 2 displays the exact number of observations we actually generated for

[10]In principle, the tie-breaking privilege should be worth more in groups of four than in groups of eight. In practice, however, ties were rare.

Table 2. Number of Observations for Each Treatment

	Number of Sessions	Individuals	Groups
$n = 4$, no leader	10	960	120
$n = 4$, leader	11	1,032	132
$n = 8$, no leader	10	1,885	120
$n = 8$, leader	11	2,112	132
All Treatments	42	5,989	504

each treatment. Most of this paper concentrates on our new findings on the behavior of ersatz monetary policy committees—the 504 experimental observations listed in the far right column of table 2.

3. Groups versus Individuals: A Replication

We turn first, and very briefly, to the 5,989 observations on individual performance and, especially, to the comparisons between groups and individuals that were the focus of Blinder and Morgan (2005). The results here are easy to summarize: for the most part, our new results with the Berkeley sample replicate what we found earlier with the Princeton sample.[11]

To begin with, we found in our Princeton experiment that groups (which were all of size five) turned in better average performances than did individuals. Specifically, the average group score was 88.3, while the average individual score was 85.3. The difference of 3 points, or 3.5 percent, was highly significant. If we merge all four of our group treatments in the Berkeley experiment, the average group score is 86.6 versus an average individual score of 81.1. Again, groups do better, but here their advantage is 5.5 points, or 6.8 percent—almost twice as large as in the Princeton experiment. This performance gap is also highly significant ($t = 11.2$).

The following regression confirms that this quantitative (but not qualitative) difference between the two experimental results is significant. Clustering by session to produce robust standard errors yields

[11] However, the Princeton and Berkeley samples have different statistical properties, including both first and second moments, which is why we abandoned our original idea of simply merging the two samples.

the following linear regression, with standard errors in parentheses and absolute values of t-ratios under that:[12]

$$\begin{array}{llll} S_i = 85.27 \; + \; 3.02\, GP_i - 4.18\, BERK_i + 2.50\big(GP_i \; {}^*BERK_i\big) & & & \\ \quad (0.37) \qquad (0.57) \qquad (0.55) \qquad (0.75) & & & \\ \quad t = 231.8 \;\; t = 5.4 \qquad t = 7.6 \qquad t = 3.4 & & & \\ \qquad\qquad\qquad\qquad R^2 = 0.027 \qquad N = 8,893 & & & (4) \end{array}$$

where *GP* and *BERK* are dummy variables associated with observations that occurred when the game was played as a *group* and by *Berkeley* students, respectively. The coefficient estimates, all of which are significant at the 1 percent level, reveal that Berkeley students perform worse than Princeton students but improve more from group interaction. We do not have a ready explanation for this difference, but we do note that Lombardelli, Proudman, and Talbot (2005, 194) found that weaker players improved more over the course of their entire experiment—spanning both group and individual play.

This finding, plus some others to be mentioned below, suggests a systematic pattern: weaker players may gain more from exposure to group play. To investigate this phenomenon a bit further, we disaggregated both our Berkeley and Princeton samples to see whether the *increase* in scores from part 1 (individual play) to part 2 (group play) correlated *negatively* with the part 1 scores. That is, do weaker players gain more from working in groups? To assess ability, we employ a natural, high-quality control for the skill of each group—namely, the average score of the group's members *prior to* group play, i.e., in part 1. We call this variable A or *Ability*. Regressing the mean score of a group over its twelve repetitions (*Gmean*) on A leads to

$$\begin{array}{lll} Gmean_i = 56.77 \; + \; 0.386\, A_i \quad R^2 = 0.320 \quad N = 351 & & \\ \qquad\quad (8.90) \quad (0.11) & & \\ \qquad\quad t = 6.38 \;\; t = 3.50 & & (5) \end{array}$$

[12]Clustering by session allows for the possibility of autocorrelation and heteroskedasticity for observations generated in a given session (i.e., by the same group of individuals). See White (1980).

Notice that the coefficient on the average individual score is considerably below unity, implying that $Gmean - A$, the improvement in group play, is decreasing in A. Thus, consistent with the findings of Lombardelli, Proudman, and Talbot (2005), we find that weaker players improve more than do stronger players from group interaction.

The next question pertains to the decision-making lag. How much time elapses, on average, between the shock and the monetary policy reaction to it? And do groups display systematically longer lags than individuals? Remember, the most surprising result from our original Princeton experiment was that groups were *not* slower; in fact, they were slightly faster, though not significantly so. Approximately the same is true in our Berkeley experiment. The mean lags before the *first* interest rate change are essentially identical (roughly 3.3 "quarters") in both group and individual play.

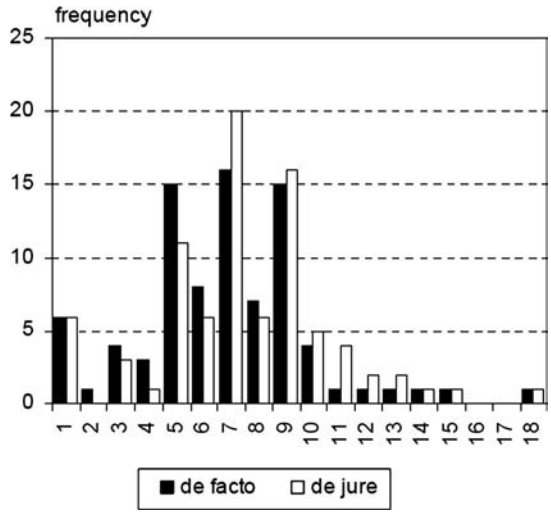
To investigate this question, we create the dependent variable *Lag*, defined as the number of quarters that elapse between the shock (the increase or decrease in G) and the committee's *first* interest rate change. Regression (6) estimates the same specification as (4), but with *Lag* replacing S as the dependent variable:

$$\begin{array}{lllll} Lag_i = 2.45 & - & 0.15\ GP_i + 0.75\ BERK_i + 0.12\ GP_i\ ^*\ BERK_i & & \\ \quad (0.23) & & (0.21) & (0.28) & (0.30) \\ \quad t = 10.7 & & t = 0.7 & t = 2.7 & t = 0.4 \\ & & & & R^2 = 0.007\ \ N = 8{,}893 \end{array} \tag{6}$$

This regression shows that groups take about the same amount of time as individuals to reach a decision, as we found before. (The F-test for omitting the two GP variables has a p-value of 0.69) It also shows that Berkeley students playing as individuals move more slowly (by approximately 0.75 "quarters") than Princeton students.

This demonstrated ability to replicate our earlier results enhances confidence in the experimental apparatus. So we turn now to the two new questions, which pertain to group size and leadership. Since the two issues are largely orthogonal, we treat them separately at first. Later (cf. table 4), we will show that interaction effects between group size and leadership are negligible and statistically insignificant.

Figure 1. Distribution of MPC Size in the Sample



Source: Erhart and Vasquez-Paz (2007)

4. Are Larger Groups More Effective Than Smaller Groups?

The title of our 2005 paper asked metaphorically, are two heads better than one? We now ask—literally—whether eight heads are better than four; i.e., do smaller ($n = 4$) or larger ($n = 8$) groups perform better in conducting simulated monetary policy?

As an empirical matter, most real-world MPCs cluster in the five- to ten-member range, with some smaller and some larger.[13] The most recent study of committee size, by Erhart and Vasquez-Paz (2007), finds that both the mean and median sizes of committees are around seven members. Figure 1, which is taken from that paper, also illustrates that the distribution of committee size is asymmetric—with a small but long right-hand tail.[14] In addition, it

[13]See Mahadeva and Sterne (2000).

[14]Erhart and Vasquez-Paz (2007) distinguish between de facto and de jure size, which do not always match up. Figure 1 shows both.

can be seen that committees with odd numbers of members are far more common than committees with even numbers. So our larger committees are somewhat typical of real-world MPCs, while our smaller committees are clearly on the small side. But does group size matter at all?

To focus on size effects, we begin by pooling the data from sessions with and without designated leaders—a pooling that our subsequent results say is legitimate. Initially, we do not control for the skill levels of the members of the group either. Simply regressing the average game score (the variable S defined above) for each of the 504 group observations on a dummy for the size of the group, and clustering by session to produce robust standard errors, yields the simple linear regression shown in column 1 of table 3, with standard errors in parentheses and asterisks indicating significance levels. The "large-group dummy" is defined to be 1 for eight-person groups and 0 for four-person groups. Thus, the regression suggests a small positive effect of larger group size—a score 2.3 points higher for the larger groups—which is significant if you are not too fussy about significance levels (the p-value is 0.067).

However, larger groups might simply have drawn, on average, more highly skilled individuals than did smaller groups. So it seems advisable to control for the abilities of the various members of the group. Fortunately, we have a natural, high-quality control for ability: the average score of all the members of the group *prior to* their exposure to group play, i.e., in part 1 of the experiment. This is the variable A introduced above, and we use both it and its square as controls for skill in the column 2 regression. Notice the huge jump in R^2—the *Ability* variable has high explanatory power.[15]

Column 2 reveals that controlling for differences in the average ability of members of the larger groups reduces the estimated difference in the performance of large versus small groups by over 40 percent—to just 1.3 points. However, even after accounting for the ability of group members, larger groups perform significantly better (p-value = 0.08) than smaller groups.

[15]When the same regression is estimated by ordinary least squares, the coefficients are almost identical, but the standard errors are roughly half of those in column 2—indicating that clustering matters.

Table 3. Regression Results on Group Size

Dependent Variable→	**(1) Score**	**(2) Score**	**(3) Score**	**(4) Score**	**(5) *Lag***	**(6) *Correct***	**(7) *Frequency***
Large-Group Dummy	2.280*	1.292*	1.033	1.031	−0.021	−0.011	−0.269*
	(1.211)	(0.723)	(0.655)	(0.657)	(0.429)	(0.038)	(0.150)
Ability		9.628***	7.027***	7.077***	−2.331**	0.006	−0.133
		(3.276)	(2.421)	(2.629)	(0.912)	(0.114)	(0.366)
$Ability^2$		−0.060***	−0.044**	−0.044**	0.014**	0.000	0.001
		(0.021)	(0.016)	(0.017)	(0.006)	(0.001)	(0.002)
Best			2.023	1.981			
			(1.857)	(1.902)			
$Best^2$			−0.010	−0.010			
			(0.012)	(0.012)			
Group Standard Deviation				0.020			
				(0.156)			
Constant	85.479***	−300.528**	−293.160***	−293.370***	97.332***	0.437**	6.068
	(1.063)	(124.108)	(85.630)	(86.630)	(33.718)	(4.256)	(13.616)
No. of Observations	504	504	504	504	504	504	504
R^2	0.03	0.24	0.26	0.26	0.07	0.01	0.03

Notes: Robust standard errors clustered by session are in parentheses. *, **, and *** denote significance at the 10 percent, 5 percent, and 1 percent level, respectively. "Large-Group Dummy" equals 1 if the group contained eight members. "*Ability*" is the average score of all members of a group during part 1 of the experiments. "*Best*" is the highest average score attained by a member of a group during part 1. "Group Standard Deviation" is the standard deviation of average part 1 scores.

The estimated quadratic in *Ability*, by the way, carries an interesting and surprising implication: that the contribution of individual ability to group performance peaks at $A = 80.7$ points, which is only a few points above the average part 1 score of 77.4 points. After that, too many good cooks seem to spoil the broth.

The negative slope beyond $A = 80.7$ is, however, an artifact of the inflexible quadratic functional form. When we estimate instead a freer functional form (such as a spline) that allows the relationship between S and A to flatten out beyond, say, $A = 80$, we get essentially a zero (rather than a negative) slope for high values of A. Still, it is surprising that groups reap no further rewards from the abilities of their members once A exceeds a fairly modest level (approximately 80). But this is a robust finding that survives experiments with several functional forms.[16]

Let us now return to why larger groups perform (slightly) better than smaller groups. One possibility is that a group's decisions are dominated by its most skilled player.[17] Larger groups will, on average, have better "best players" than smaller groups simply because the first-order statistic for skill will, on average, be higher when $n = 8$ than when $n = 4$. To see whether that factor might be empirically important in these data, we added both the average score of the group's best player (*Best*) and its square in the regression to get the regression reported in column 3. We see that the effect of larger group size is reduced by about 20 percent, and it is now no longer significant at even the 10 percent level ($p = 0.12$).

The explanatory power of the *Best* variables is modest, however. Neither *Best* nor $Best^2$ is statistically significant on its own, and the estimated coefficients are small compared with those of the A variables. Moreover, adding *Best* and $Best^2$ raises R^2 by only 0.026.[18] However, an F-test of the joint hypothesis that the coefficients on

[16]The surprising thing is not that there are diminishing returns to group size, which can be rationalized in many ways, but that marginal returns seem to get to zero so quickly.

[17]Several colleagues *assured* us that this would be the case in our first experiment. But we tested and rejected the hypothesis in Blinder and Morgan (2005).

[18]Surprisingly, the individual score of the *second-best* player turns out to have more explanatory power for the group's performance. We have no ready explanation for this finding and treat it as a fluke. Regardless, the results on group size are not qualitatively affected under this alternative specification.

both variables are 0 strongly rejects that hypothesis ($F = 30.9$, $p = 0.00$).[19] Thus, the evidence suggests that the fuller specification (column 3) is preferred, but that the influence of the best player is modest—a point to which we shall return in considering the effects of leadership.

Next, we ask whether heterogeneity of the members of the group, as measured by skill differences across players, improves group performance. We measure heterogeneity by introducing a new variable in column 4: the standard deviation of the average part 1 scores obtained by the (four or eight) members of the group.[20] Comparing columns 3 and 4 shows that adding this variable has essentially no effect on the regression. Thus heterogeneity does not seem to matter.

4.1 How Do Larger Groups Outperform Smaller Groups?

Having shown that larger groups (barely) outperform smaller groups, the next question is, how do they do it? To determine whether a shorter or longer decision-making lag is the source of the advantage for large groups, we regress the variable *Lag* defined above on the dummy for groups of size eight and the ability controls mentioned above, clustering by session as usual. The result is the regression in column 5 of table 3, which indicates no difference between the two group sizes in terms of speed of decision making. (The p-value of the coefficient of the dummy is 0.58.) Differences in ability are again significant, with groups composed of more skilled players tending to decide more quickly—but only until A reaches 81.2. Moreover, note the low R^2 in this regression, which indicates that neither group size nor ability explains much of the variation in lag times.

Next, we turn to *accuracy* rather than *speed*. For the regression in column 6, we define a new left-hand variable, *Correct*, which is

[19]This looks like the classic symptoms of extreme multicollinearity, but in fact the correlation between A (the group average) and *Best* is only 0.67. Replacing A—which, of course, includes *Best*—with the median does not reduce the multicollinearity at all (the correlation between the median and *Best* is also 0.67), and it generally produces worse-fitting regressions. For these reasons, we stick with the mean, rather than the median, in what follows.

[20]This is an admittedly narrow concept of heterogeneity. But, other than the sex composition of the group (which did not matter), it is the only measure of heterogeneity we have.

equal to 1 if the group's initial interest rate move is in the correct direction—i.e., if a rise in G is followed by a monetary tightening, or a decline in G is followed by a monetary easing—and equal to 0 otherwise. Do larger groups derive their advantage by being more accurate, in this sense?[21] Apparently not. The group-size dummy again shows no difference between groups of size four and size eight. It is interesting to note that, now, the average ability of the members of the group is also of no use in predicting the group's success—a surprising finding.

Having failed so far, we turn to one last performance metric: the frequency of interest rate changes. Remember that each change in the rate of interest costs the group a 10-point charge. So it is possible that larger groups do better because they "fiddle around" less with interest rates. To find out, we define a new left-hand variable, *Frequency*, which measures the number of rate changes a group makes over the course of a twenty-quarter game. Since interest rate changes are costly, it pays for groups to economize on them. The simple regression in column 7 reveals a modest effect of group interaction in producing more "patient" decision making. And, strikingly, the *Ability* variable seems to have little to do with the frequency of rate changes.

Here at last we find a partial answer to the question of why larger groups perform slightly better: they average 0.27 fewer interest rate changes per game (with a p-value of 0.08). Since only about 2.25 changes are made on average, this is a meaningful difference.

To summarize this investigation, larger groups take about as much time (measured in terms of data) and are about as accurate in their decisions as smaller groups. However, they make slightly fewer interest rate changes overall and, in this (limited) sense, are slightly more "stodgy" decision makers than individuals. This slightly more patient behavior, in turn, produces a systematic, though quite modest, performance advantage over small groups.

[21] Of course, since *Correct* is binary, a linear probability specification may not be appropriate. As an alternative, we could have performed a probit regression at the cost of not being able to cluster standard errors. The results from probit regressions are qualitatively and quantitatively similar to the linear probability specifications reported here.

Why might larger groups do slightly better? In this environment of pervasive uncertainty, each member of a group is likely to carry in a somewhat different notion of how the model works from his or her own experience with individual play—and thus, in particular, a different notion of how often to change interest rates. Group play allows members to pool the wisdom gained from their individual experiences. If pooling offers gains, but the gains are subject to diminishing returns, we might find groups of eight outperforming groups of four.

But then why are the gains from larger group size so small? One possibility might be that the optimal committee size is, say, $n = 6$. In that case, committees of four (too small) and eight (too large) might be almost equally suboptimal.[22] Alternatively, it could be that $n = 4$ and $n = 8$ are simply too close together, and that experimenting with, say, $n = 12$ or more might have produced larger differences. Finally, it is worth noting that our committees are all symmetric—everyone does exactly the same thing. By contrast, many real-world MPCs allow (or require) their members to specialize in certain tasks. Diminishing returns presumably sets in more slowly in such specialized committees than in symmetric committees.

5. Does Leadership Enhance Group Performance?

As noted in the introduction, virtually all decision-making groups in the real world, and certainly all MPCs, have well-defined leaders—e.g., the chairman of a committee. To an economist, or to a Darwinian evolutionist for that matter, this observation creates a strong presumption that leadership must be productive—for why else would it be so ubiquitous? But, as we show now, our experimental findings say otherwise: surprisingly, groups with designated leaders do *not* outperform groups without leaders.

We start table 4, as we did table 3, with a simple regression comparing the scores of groups with and without leaders—ignoring, for the moment, both average ability and group size. The leader dummy, defined to be 1 if the group has a designated leader and 0 otherwise, actually gets a *negative* (though insignificant) coefficient

[22]This possibility was suggested to us by Petra Geraats.

Table 4. Regression Results on Leadership

Dependent Variable→	**(1) Score**	**(2) Score**	**(3) Score**	**(4) Score**	**(5) *Lag***	**(6) *Correct***	**(7) *Frequency***	**(8) Score**
Leader Dummy	−0.832	−0.160			−0.287	−0.025	0.154	−0.718
	(1.225)	(0.742)			(0.415)	(0.033)	(0.152)	(1.098)
Ability		10.300***	12.257*	17.820***	−2.377***	0.009	−0.259	9.430***
		(3.515)	(6.098)	(4.138)	(0.822)	(0.102)	(0.347)	(3.182)
$Ability^2$		−0.064***	−0.078*	−0.114***	0.015**	0.000	0.002	−0.058***
		(0.023)	(0.041)	(0.028)	(0.006)	(0.001)	(0.002)	(0.021)
Best			−0.384					
			(2.697)					
$Best^2$			0.005					
			(0.017)					
Female				−1.164				
				(1.100)				
Large-Group Dummy								0.769
								(0.837)
Large Group × Leader								1.045
								(1.439)
Constant	87.054***	−325.405**	−393.587*	−609.677***	99.346***	0.352	10.582	−292.001**
	(0.613)	(133.642)	(202.219)	(153.455)	(30.251)	(3.825)	(13.029)	(120.957)
No. of Observations	504	504	264	252	504	504	504	504
R^2	0.01	0.23	0.32	0.32	0.07	0.01	0.02	0.03

Notes: Robust standard errors clustered by session are in parentheses. *, **, and *** denote significance at the 10 percent, 5 percent, and 1 percent level, respectively. "Leader Dummy" equals 1 if the group had a designated leader. "*Ability*" is the average score of all members of a group during part 1 of the experiment. "*Best*" is the highest average score attained by a member of a group during part 1. "*Female*" equals 1 if the leader was a female. "Large-Group Dummy" equals 1 if the group contained eight members.

in column 1. Once we control for ability in column 2, this small coefficient drops to virtually zero, and the effect of ability on group performance resembles that in table 3—a quadratic in A that peaks at 80.4. Thus the counterintuitive finding is that leadership does *not* affect group performance. We proceed now to try to overturn this surprising non-result.

One obvious explanation might be that our designated leaders achieve their high scores during part 1 purely by chance, and thus are really no more able than the others. This possibility is easily dismissed, however, by looking at scores in part 3 of the game—when subjects play again as individuals. Across all individuals who participated in the sessions with designated leaders, the correlation between their part 1 scores and their part 3 scores is 0.45, indicating substantial and durable individual effects. Thus it is not just luck; the leaders really are better.

One interesting question to ask, once again, is whether the group's score is driven more by the skill of the average member or by the skill of the leader. To address this question, we restrict our attention in column 3 to sessions with designated leaders (thus reducing the sample size to 264) and add the previously defined variables *Best* and $Best^2$ to the regression. Remember that *Best* is the average score of the highest-scoring individual in part 1—and thus the score of the designated the leader in part 2.

Interestingly, the average skill of the group's members ("*Ability*") is a much better predictor of performance than the skill of the leader ("*Best*"). To see this formally, we ran F-tests to determine the effect of omitting the two *Ability* variables versus omitting the two *Best* variables from the regression. For the *Ability* variables, the F-statistic is 8.7 ($p = 0.00$) whereas for the *Best* variables, the F-statistic is only 3.2 ($p = 0.06$). The comparative weakness of the *Best* variables illuminates the puzzling absence of leadership effects on performance: while the leader is the best player, he or she seems incapable of improving the performance of the group.[23]

We next ask whether leadership effects on group performance differ by the gender of the leader by adding a dummy variable *Female*

[23]The inverted quadratic in *Best* seen in column 3 looks peculiar, but it is upward sloping in the relevant range. Given the imprecision of the estimates of these coefficients, one shouldn't make much of this result.

to the regression in column 4. Again, we restrict our attention to sessions with designated leaders.[24] While the estimated coefficient for female leaders is negative, it does not come close to statistical significance. Thus, women do neither better nor worse as leaders.[25]

So leaders seem to have no discernible effect on their group's *score*. But do they influence the group's *strategy*? To examine this question, we look next at the dependent variable *Lag* defined earlier. Column 5 shows that leadership does not influence the speed of reaction significantly. While the coefficient of the leader dummy is negative, it is insignificant.

What about leadership effects on the likelihood of moving in the correct direction on the first interest rate change? The column 6 regression also shows essentially no effect.

Finally, we turn to the frequency of rate changes. Do groups with designated leaders change interest rates more (or less) frequently? The answer is (weakly) more frequently, as the regression in column 7 shows. But, again, the effect does not come close to statistical significance.

To this point, we have looked for leadership effects on the (tacit) assumption that they are the same in large ($n = 8$) and small ($n = 4$) groups. Similarly, in the previous section we examined the effects of group size while maintaining the hypothesis that size effects are the same with and without leaders. To test for possible interaction effects, the last regression in table 4 includes dummies for both group size and leadership, allowing an interaction between the two.

Column 8 reveals that the interaction effect is totally insignificant (p-value $= 0.47$). Still, if the positive coefficient is taken at face value, the regression suggests a small *negative* effect of leadership in smaller groups and a small *positive* effect in larger groups.

A fair summary so far would be to say that you need a magnifying glass—and you must ignore statistical significance—to see any effects of leadership on group performance. The main message, surprisingly, is that leadership does not seem to matter.

[24] A leader in one of the eight-person sessions refused to identify his or her gender, which reduced the number of observations to 252.

[25] They are also neither better nor worse as followers. The sex composition of the group does not help explain the group's performance.

Table 5. Improvements from Individual to Group Play, by Treatment

(1) Treatment	(2) Part 1 Mean Score (Individual Play)	(3) Part 2 Mean Score (Group Play)	(4) Difference
$n = 4$, no leader	78.4	87.1	8.7 (11.1%)
$n = 4$, leader	75.5	84.1	8.6 (11.4%)
$n = 8$, no leader	76.8	87.1	10.3 (13.4%)
$n = 8$, leader	78.2	88.4	10.2 (13.0%)

One other place to look for leadership effects is in how much people learn from their experience playing as a group. In our Princeton and Berkeley experiments, we found significant improvements in performance when individuals came together to play as groups. Could it be that the learning that takes place in group play is greater when the group has a designated leader?

Table 5 displays the *improvements* in score from part 1 (individual play) to part 2 (group play) separately for each of the four experimental treatments. Column 4 shows no support for the idea that group interactions help subjects more when there is a designated leader.

To assess statistical significance, we examine the dependent variable $DIFF_i$ suggested by table 5: the average score of a given subject in group play (part 2 of the game) *minus* that same individual's average score while playing as an individual in part 1. Table 5 suggests that improvements are slightly higher with larger groups but are independent of leadership. Thus, we include as right-hand variables dummies for both group size and whether the group is led or not. As usual, we cluster by session to obtain

$$\begin{array}{lllll} DIFF_i = 8.71 & + \ 0.03\ LED_i + 1.46\ D8_i & R^2 = 0.005 & N = 250 \\ \quad (0.83) & \quad (0.99) \quad\ \ (0.99) & & \\ \quad t = 10.5 & \quad t = 0.03 \quad\ \ t = 1.5 & & \end{array} \tag{7}$$

where LED is the leader dummy and $D8$ is the large-group dummy.

Table 6. Improvements from Part 1 to Part 3, by Treatment

(1) Treatment	(2) Part 1 Mean Score (Individual Play)	(3) Part 3 Mean Score (Individual Play)	(4) Difference
$n = 4$, no leader	78.4	83.2	4.8 (6.1%)
$n = 4$, leader	75.5	85.2	9.7 (12.8%)
$n = 8$, no leader	76.8	85.1	8.3 (10.8%)
$n = 8$, leader	78.2	84.9	8.7 (8.6%)

This regression shows that leadership has no effect on the improvement between individual and group play. On the other hand, participation in larger groups does improve upon individual performance slightly more than participation in smaller groups; however, the result does not quite rise to the level of statistical significance ($p = 0.15$).

One final question about leadership and learning can be raised. We found in both experiments that scores typically improve quite a bit when subjects move from individual play to group play (from part 1 to part 2) but then fall back somewhat when they return to individual play (from part 2 to part 3). The change in an individual's performance from part 1 to part 3 can therefore be used as an indicator of what might be called the "durable learning" that emerges from experience with group play. Is this learning greater when the group has a designated leader than when it does not?

Table 6 suggests that the answer is no. The subjects learn more from group play with a designated leader when $n = 4$, but less when $n = 8$. Notice, by the way, that the largest improvement in table 6 comes in the $\{n = 4, \text{leader}\}$ groups—the treatment that, by chance, got the weakest players.

The statistical significance of this result can be appraised by regressing the dependent variable $POSTDIFF_i$, defined as the difference between the average score of a given subject in part 3 of the game less that same individual's average score in part 1, on dummy

variables for leadership and size. Clustering by session as usual, the result is

$$\begin{array}{lllll} POSTDIFF_i = 7.38 + & 0.41\,LED_i - & 0.18\,D8_i & R^2 = 0.00 & N = 250 \\ \quad (1.13) & \quad (1.21) & \quad (1.21) & & \\ \quad t = 6.5 & \quad t = 0.3 & \quad t = 0.2 & & \end{array} \tag{8}$$

This regression shows that neither group size nor leadership affects the durable performance gains that arise from exposure to group play.

In sum, there is no evidence from our experiment of superior (or even faster) performance by groups with designated leaders versus groups without. Overall, the most prudent conclusion appears to be that groups with designated leaders perform no differently than groups without leaders. This is a surprising finding, to say the least. Should we believe it? Maybe, but maybe not.

5.1 Why No Leadership Effects?

First, in defense of our experimental design, remember that we do *not* choose the leaders randomly or arbitrarily. Rather, each designated leader *earns* his or her position by superior performance *in the very task that the group will perform.* This principle for selecting leaders, we believe, imbues them with a certain legitimacy—as is normally the case in real-world groups. A second element of realism derives from the reward structure. Doubling the leader's reward in group play gives him or her a greater stake in the outcome—just as leaders of real-world groups normally have greater stakes in the outcome than other members do. For example, history will appraise the performance of the "Bernanke Fed" and the "Roberts Court." The names of most of the other members will be lost to history.

Second, however, while giving the leader the tie-breaking vote allows him or her to influence the group's decisions *in principle*, it may not do so *in practice.* For example, we found in Blinder and Morgan (2005) that there was no difference in either the quality or speed of group decision making when groups made decisions unanimously rather than by majority rule. And, as noted earlier, tie votes were rare in this experiment.

Third, and in a similar vein, we are able to test only for differences between groups with and without an *officially designated* leader; we have no independent measurement of how *effective* leadership is. Thus, some of our putative leaders may actually be quite passive, while strong leadership could emerge spontaneously in some of the groups without a designated leader.

Fourth, it should be noted that the task in our experimental setup is what psychologists call intellective (figuring something out) rather than, say, judgmental or moral (deciding what's right and wrong). So the surprising conclusion that leadership in groups has no apparent benefits should, at the very least, be limited to such intellective tasks. As Fiedler and Gibson (2001, 171) pointed out, "Extensive empirical evidence has shown that a leader's intellectual ability or experience does not guarantee good [group] performance." That said, making monetary policy decisions is, for the most part, an intellective task.

Fifth, however, there is never any disagreement among members of our ersatz MPCs over what the group's objectives are (including the relative weights). Every player tries to maximize exactly the same function. By contrast, there is potential for disagreement over the central bank's objectives and/or weights on at least some real-world MPCs (e.g., the FOMC)—which might allow more scope for effective leadership. In fact, this raises a broader issue. Our student subjects are arguably a more homogeneous group than at least some MPCs, to which people of diverse backgrounds are deliberately appointed.

Sixth, and related, our committees deal only with "normal" monetary policy decisions. It is certainly possible that greater scope for leadership might emerge if our experimental subjects were faced with crises, such as the ones the Federal Reserve and the ECB have been confronted with in 2007 and 2008.

Finally, and perhaps most important, our narrow experimental concept of leadership—leading the discussion, reporting the group's decision, and breaking a tie if necessary—does not correspond to the common meaning of "leadership" as expressed, e.g., in the admittedly chauvinistic statement "He's a leader of men." Our experimental leaders do not lead in the sense that a military officer leads a platoon, a politician leads a party, or an executive leads a business. Brown, Scott, and Lewis (2004) classified leaders as

"transformational" and "transactional," the latter meaning motivating subordinates with rewards. Our experimental leaders were neither.

We thought about trying to select group leaders by what might loosely be described as "leadership qualities" but quickly abandoned the idea as being too subjective and too difficult. We think this decision was the right one. But, in interpreting the experimental results, it is important to remember that our leaders are selected, on average, for their "smarts," not for their "leadership qualities." There is no reason to think that the cognitive ability we use to select group leaders correlates highly with the traits that are associated with leadership in the real world, such as verbal dexterity, aggressiveness, an extroverted personality, a trustworthy affect, good looks, and height. However, we certainly hope (and believe) that cognitive ability is a relevant consideration in the selection of real-world central bank heads.

Similarly, it seems plausible that true—as opposed to putative—leadership in groups may need to emerge slowly over time, as the leader demonstrates good performance and as other members grow to respect his or her judgment, acumen, and group-management skills. A one-time, ninety-minute laboratory experiment leaves no scope for that sort of leadership to emerge.

Thus we certainly do not believe that our experimental results provide the last word on leadership effects. We offer them as something closer to the first word, and invite other researchers to pick up the challenge.

6. Conclusions

In this paper, we replicate earlier findings from Blinder and Morgan (2005) showing that simulated monetary policy committees make systematically better decisions than the same individuals making decisions on their own, without taking any longer to do so. This experimental evidence supports the observed worldwide trend toward making monetary policy decisions by committees, rather than by lone-wolf central bankers. We also find several suggestive shreds of evidence that the margin of superiority of groups over individuals is greater when the individuals are of lower ability.

But the more novel findings of this paper pertain to groups that differ in terms of size and leadership. We find some weak evidence that larger groups (in our case, $n = 8$) outperform smaller groups ($n = 4$), mainly because larger groups seem better able to resist the temptation to "fiddle" with interest rates too much. But these differences are small, and many are not statistically significant. So, in terms of institutional design, it is not clear whether larger or smaller MPCs are to be recommended. (Remember that $n = 7$ is the mean and modal size of real-world MPCs.)

Our most surprising and important result, at least to us, is that ersatz MPCs do *not* perform any better when they have a designated leader than when they do not—even though every real-world MPC has a clear (and sometimes dominant) leader, and even though our designated leaders were chosen on the basis of their skill in making monetary policy. We caution that we would not apply this finding beyond the realm of intellective tasks—e.g., we do not recommend that army platoons venture out without a commanding officer!

But that said, there are probably many more intellective than combative tasks in the economic world, certainly including monetary policy. For example, promotions in business are often based on superior performance on metrics that are basically intellective. So this finding, if verified by other work, is potentially of wide applicability. In terms of the taxonomy of MPCs emphasized by Blinder (2004), our results suggest that an *individualistic* committee, where the leader is only modestly more important than the other members, may function just as well as a *collegial* committee, where the role of the leader is more pronounced.

Appendix. The Instructions

Note: Portions of the instructions read only during sessions with leaders are enclosed in brackets.

In this experiment, you make decisions on monetary policy for a simulated economy, much like the Federal Reserve does for the United States. At first, you will make the decisions on your own; later, we will bring you all together to make decisions as a group. [At that point, one of you will be designated the leader of the group, as I will explain shortly.]

We have programmed into each computer a simple model economy that generates values of unemployment and inflation, period by period, for twenty periods. Think of each period as a calendar quarter, so the game represents five years. Each quarterly value of unemployment and inflation depends on the interest rates you choose and on some random influences that are beyond your control. Every machine has exactly the same model of the economy, but each of you will get different random drawings and so will have different experiences.

Your goal is to keep unemployment as close to 5 percent, and inflation as close to 2 percent, as you can—quarter by quarter. As you can see from the top line on the screen, we start you off with an interest rate of 7 percent in period 1. Initially, unemployment and inflation differ slightly from the targets of 5 percent and 2 percent because of the random influences I just mentioned. Starting with period 2, you must choose the interest rate.

Raising the interest rate will *increase* unemployment and *decrease* inflation. But the effects are delayed—neither unemployment nor inflation responds immediately. Similarly, lowering the interest rate will decrease unemployment and increase inflation. But, once again, the effects are delayed.

The computer determines your score for each period as follows. Hitting 5 percent unemployment and 2 percent inflation exactly earns you a perfect score of 100 points. For each tenth-of-a-point by which you miss each target, you lose a point on your score. Direction doesn't matter; you lose the same amount for being too high as for being too low. Thus, for example, 5.8 percent unemployment and 1.5 percent inflation will net you a score of 100 *minus* 8 points for missing the unemployment target by eight-tenths *minus* 5 points for missing the inflation target by five-tenths, or 87 points. Similarly, 3.5 percent unemployment and 3 percent inflation will net you $100 - 15 - 10$, or 75 points. If you look at the top line of the display, you can see that the initial unemployment rate of 5 percent and inflation rate of 1.9 percent yields a score of 99.

Finally, there is a cost of 10 points each time you change the interest rate. The 10 points will be deducted from that period's score.

Are there any questions about the scoring system?

As you progress through the experiment, accumulating points both in individual and in group play, the computer will keep track of your cumulative *average* score on the 1–100 scale. At the end of the session, your cumulative average score will be translated into money at the rate of *25¢ per point*, and you will be paid your winnings by check. Thus a theoretical perfect score of 100 would net you $25, an average score of 80 would give you 80 percent of $25, or $20, etc. You are guaranteed a minimum of $15, no matter how badly you do.

[When you play as individuals, everyone is treated the same. But when we bring you together to play as a group, one of you will serve as the group's *leader*. The leader will be the one who scored the highest in individual play, and he or she will receive *twice* as many points during group play.]

The game works as follows. You can move the interest rate up or down, in increments of 1 percentage point, by moving the slide bar on the left-hand side of the screen, or by clicking on the up or down buttons. Try that now to see how it works. When you have selected the interest rate you want, click on the button marked "Click to Set Rate." Do that now to see how it works. The computer has recorded your choice, drawn the random numbers I mentioned earlier, and calculated that period's unemployment, inflation, and score.

There is one final, important aspect to the game. At a time period selected at random, but equally likely to be any of the first ten periods, aggregate demand will *either* increase *or* decrease. You will not be told *when* this happens nor in *which direction*.

If aggregate demand *increases*, that tends to push unemployment down and, with a lag, inflation up. If aggregate demand *decreases*, that tends to push unemployment up and, with a lag, inflation down. The essence of your job is to figure out when and how to adjust interest rates in order to keep unemployment as close to 5 percent, and inflation as close to 2 percent, as possible.

Remember, the change in aggregate demand comes at a randomly selected time within the first ten periods, and we will not tell you whether demand has gone up or down. And each interest rate change will cost you 10 points in the period you make it.

Are there any questions?

Please sign the consent form located in the folder next to your computer.

This will all be simpler once you've practiced on the apparatus a bit. You can do so now, and your scores will just be displayed for your information; they will not be recorded or counted. You can practice for about five minutes to develop some familiarity with how the game works. During this practice time, feel free to ask any questions you wish.

OK, it's time to start the game for real now.

In the first part of the experiment, you will play the monetary policy game twelve times *by yourselves*. After you have played the game twelve times, the computer will prevent you from going on. You may not communicate with any other player, and the points you earn will be your own.

[As I mentioned, the student who earns the highest score in this part of the experiment will be the leader of the group in the next part.]

Please start now by clicking the continue button, and proceed at your own pace.

Now please gather around the projection screen to play the same game as a group. [The highest scoring player will be the leader.]

In this part of the experiment, you will play exactly the same game twelve times. The rules are the same except that decisions are now made *by majority rule*. [In case of a tie, the leader will cast the tie-breaking vote. The leader will control the mouse.] You may communicate freely with each other, as much as and in any way you wish. While playing as a group, each of you will receive the group's score [except the leader, who will earn twice as many points]. At the conclusion of group play, we will show you how your performance compared with the top scores achieved when we ran this game at Princeton. Any questions?

Please begin.

OK. Now please return to your individual seats and, once again, play twelve rounds of the game by yourselves. Communication with

other players is not allowed. The computer will again stop you after twelve rounds.

Please begin.

OK. The experiment is now over. Thank you for participating.

References

Blades, J. W. 1973. "Influence of Intelligence." In *The Influence of Intelligence, Task Ability, and Motivation on Group Performance*, ed. J. W. Blades and F. E. Fiedler, 76–78. University of Washington, Seattle: Organizational Research Technical Report.

Blinder, A. S. 2004. *The Quiet Revolution: Central Banking Goes Modern*. New Haven, CT: Yale University Press.

Blinder, A. S., and J. Morgan. 2005. "Are Two Heads Better Than One? Monetary Policy by Committee." *Journal of Money, Credit, and Banking* 37 (5): 789–812.

Brown, D., K. Scott, and H. Lewis. 2004. "Information Processing and Leadership." In *The Nature of Leadership*, ed. R. Sternberg et al. New York, NY: Sage Publications.

Chappell, H. W., Jr., R. R. McGregor, and T. Vermilyea. 2005. *Committee Decisions on Monetary Policy*. Cambridge, MA: MIT Press.

Edmondson, A. 1999. "Psychological Safety and Learning Behavior in Work Teams." *Administrative Science Quarterly* 44 (4): 350–83.

Erhart, S., and J. L. Vasquez-Paz. 2007. "Optimal Monetary Policy Committee Size: Theory and Cross-Country Evidence." Paper presented at Norges Bank conference.

Fiedler, F. E., and F. W. Gibson. 2001. "Determinants of Effective Utilization of Leader Abilities." In *Concepts for Air Force Leadership*, ed. R. I. Lester and A. G. Morton, 171–76. Maxwell Air Force Base, AL: Air University Press.

Guth, W., M. Vittoria Levati, M. Sutter, and E. van der Heijden. 2004. "Leadership and Cooperation in Public Goods Experiments." Discussion Paper on Strategic Interaction No. 2004-29, Max Planck Institute of Economics.

Lombardelli, C., J. Proudman, and J. Talbot. 2005. "Committees versus Individuals: An Experimental Analysis of Monetary Policy Decision Making." *International Journal of Central Banking* 1 (1): 181–205.

Mahadeva, L., and G. Sterne, ed. 2000. *Monetary Policy Frameworks in a Global Context*. New York, NY: Routledge Publishers.

Maier, N. R. F. 1970. *Problem Solving and Creativity in Individuals and Groups*. Belmont, CA: Brooks/Cole.

Rudebusch, G. 2001. "Is the Fed Too Timid? Monetary Policy in an Uncertain World." *Review of Economics and Statistics* 83 (2): 203–17.

Sibert, A. 2006. "Central Banking by Committee." *International Finance* 9 (2): 145–68.

White, H. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica* 48 (4): 817–38.