NBER WORKING PAPER SERIES

LEADERSHIP IN GROUPS: A MONETARY POLICY EXPERIMENT

Alan S. Blinder
John Morgan

Working Paper 13391
http://www.nber.org/papers/w13391

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2007

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, and seminar participants at Princeton, the International Monetary Fund, and the National Bureau of Economic Research. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

Leadership in Groups: A Monetary Policy Experiment
Alan S. Blinder and John Morgan
NBER Working Paper No. 13391
September 2007
JEL No. E52,E58

ABSTRACT

In an earlier paper (Blinder and Morgan, 2005), we created an experimental apparatus in which Princeton University students acted as ersatz central bankers, making monetary policy decisions both as individuals and in groups. In this study, we manipulate the size and leadership structure of monetary policy decisionmaking. 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. Furthermore, we find rather little difference between the performance of four-person and eight-person groups; the larger groups outperform the smaller groups by a very small margin. Finally, we successfully replicate our Princeton results, at least qualitatively: Groups perform better than individuals, and they do not require more "time" to do so.

Alan S. Blinder
Department of Economics
Princeton University
Princeton, NJ 08544-1021
and NBER
blinder@princeton.edu

John Morgan
Haas School, UC, Berkeley
545 Student Services Building, #1900
Berkeley, CA 94720-1900
morgan@haas.berkeley.edu

I. 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, Chapter 2). In an earlier paper (Blinder and Morgan, 2005), we created an experimental apparatus in which Princeton University students acted as ersatz central bankers, making monetary policy decisions both as individuals and in groups. Those experiments yielded two main findings:

1. groups made better decisions than individuals, in a sense to be made precise below;
2. groups took no longer to reach decisions than individuals did.[1]

Finding 1 was not a big surprise, given the previous literature on group versus individual decisionmaking (most of it from disciplines other than economics). But we were frankly stunned by finding 2. Like seemingly everyone, we believed that groups moved more slowly than individuals. A subsequent replication with students at the London School of Economics (Lombardelli *et al.*, 2005), verified finding 1 but did not report on finding 2.

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.

[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.

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 *et al*. (2005), following our lead, also used committees of five. But real-world monetary policy committees vary 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 nineteen, 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]

The second issue pertains to *leadership* and is the truly unique aspect of the research reported here. In both our Princeton experiment and in Lombardelli *et al.*'s replication, all members of the committee were treated equally. But every real-world monetary policy committee has a designated leader who clearly outranks the others. At the Federal Reserve, he 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 (much less so, it appears, under Ben Bernanke) was at one extreme; it

[2] The reason for choosing even-numbered groups will be made clear shortly. Our "large" groups (n=8) are still small compared to, e.g., the ECB or the Fed. This size was more or less dictated by the need to recruit large numbers of subjects. With groups of four and eight, we needed 252 subjects in all.

was what Blinder (2004, Chapter 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 *et al.* (2005, Chapter 7) estimated econometrically that Arthur Burns' views on monetary policy carried about as much weight as those of all the 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. Its 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, which strive for consensus. Some of these committees are led firmly; others are led gently.

The scholarly literature on group decisionmaking, which comes mostly from psychology and organizational behavior, gives us 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 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), despite the fact that it is often viewed as a central selection criterion for choosing leaders. Rather, the extant literature suggests that gains from group interaction may depend more on how well the leader encourages the other members of the group to contribute their opinions frankly and openly (Blades (1973), Maier (1970), Edmondson (1999)). In an interesting public goods experiment,

Guth *et al.* (2004) also 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 to have four treatments, running either 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 very 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 rather 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 II describes the experimental setup, which is in most respects exactly the same as in Blinder and Morgan (2005). Sections III and IV focus on the data generated by decisionmaking in groups, presenting new results on the effects of group size and leadership respectively. Then Section V

briefly presents results comparing group and individual performance that mostly replicate those of our Princeton experiment. Section VI summarizes the conclusions.

II. 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 that we used in the Princeton experiment—with parameters chosen to resemble the U.S. economy:

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

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

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%, and on its own four lagged values, with weights summing to one. The coefficient on the unemployment rate was 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. 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 5%. (Here i is the nominal interest rate.) But there is a lag in the relationship, so

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

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.

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 10 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 can be an equilibrium.

We begin each round of play with inflation at 2%—which is also the central bank's target rate (see below). Thus, prior to the shock (that is, when G=0), the model's steady-state equilibrium is U=5, i=7, π=2. As is apparent from the coefficients in equation (2), the shock changes the neutral real interest rate from 5% to either 6% or 4% *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 approximately 0.14, or about half the size of the G shock. This sizing decision, we found, makes the fiscal shock relatively easy to detect—but not too easy.

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

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.

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, if not greater than, 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- or backward-looking, and so on. Nor do they benefit from staff forecasts.

Furthermore, 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. Similarly, it is not easy to distinguish quickly between the permanent G shock and the transitory e and w shocks that add

[5] A copy of the instructions is available on request.

"noise" to the system—especially since subjects do not know that the G shock is permanent. Once unemployment and inflation start to "run away from you," it can be difficult to get them back on track.

Each play of the game proceeds as follows. We start the system in steady state equilibrium at the values mentioned above: G=0, i=7%, lagged U=5%, and all lags of π=2%. 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%, π=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 20 such periods. Students are told to think of each period as a quarter; so the simulation covers "five years."

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 decisionmaker 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 below that one type of decisionmaking process "takes longer" than another, we mean that more *data* (not more *minutes*) are required.

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

$$(3)\quad s_t = 100 - 10\,|\,U_t - 5\,|\, - 10\,|\,\pi_t - 2\,|\,,$$

and the score for the entire game (henceforth, S) is the (unweighted) average of s_t over the 20 quarters. We use an absolute-value function instead of the quadratic loss function that has become ubiquitous in research on monetary policy (and much else) 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, for example, 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 cents 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 20-period game, it amounts to just 0.5% 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 the Fed's reaction function from the minimization of a quadratic loss function find that

[6] The unemployment and inflation data are always rounded to the nearest tenth. So students see, e.g., 5.8%, not, say, 5.83%.
[7] To keep things simple, only integer interest rates are allowed.

they must add, say, a quadratic term in $(i_t - i_{t-1})$ to fit the data. Without that wrinkle, 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. The instructions tell them, among other things, that the person earning the highest score while playing alone in Part One of the experiment will be designated the "leader" (the term we use) of the group for Part Two—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. Scores during those practice rounds are displayed for feedback, but not recorded. At the end of the practice period, each machine is reinitialized, and each student is instructed to play 12 rounds of the game (each lasting 20 "quarters") *alone*—without communicating in any way with the other subjects. Once all the subjects have completed 12 rounds of individual play, the experimenter calls a halt to Part One of the experiment.

In Part Two, the same students gather around a single large screen to play the same game 12 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 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 One; and that student becomes the leader for Part Two.[9]

[8] See, for example, Rudebusch (2001).

Thus, the criterion for electing leaders is purely intellective: the skill of an individual at ersatz monetary policy making. Since the group will perform the identical task, this selection principle would seem a natural one.

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 normally ensures that the leader leads the discussion. Second, the leader faces higher powered incentives in the task. As just mentioned, his or her score in Part Two is *double* that of the other subjects. Third, the leader gets to break a tie vote if there is one—which is why we chose even-numbered groups.[10] While we recognize that the experimental setup still only allows limited scope for leadership, we judged that this what about all we could do in a laboratory setting with 1½ hours of experimental time. We return to this issue later.

After 12 rounds of group play, the subjects return to their individual computers for Part Three, in which they play the game another 12 times alone, with no communication with the others. For future reference, Table 1 summarizes the flow of each session.

Table 1
The Flow of the Experiment

Instructions
Practice Rounds (no scores recorded)
Part One: 12 rounds played as individuals
Part Two: 12 rounds played as a group (with or without a leader)
Part Three: 12 rounds played as individuals
Students are paid by check and leave.

[9] On average, that student scored 10.77 points higher than the others in the group during Part One of the experiment.
[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.

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 x 4 x 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 the same 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 each treatment. We concentrate on our new findings on the behavior of ersatz monetary policy committees—the 504 experimental observations listed in the righthand column of Table 2.

Table 2
Number of observations for each treatment

	Number of sessions	**Individuals**	**Groups**
n=4, no leader	10	960	120
n=4, leader	11	1032	132
n=8, no leader	10	1885	120
n=8, leader	11	2112	132
All treatments	**42**	**5989**	**504**

III. 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; that is, 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.[11] So our eight-person committees are somewhat typical of real-world MPCs while our four-person committees are 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 following linear regression, with standard errors in parentheses and the absolute values of t-ratios under that:[12]

$$(4) \quad S_i = 85.48 + 2.28\, D8_i \qquad R^2 = 0.028 \qquad N = 504 \text{ observations}$$
$$\qquad\qquad (1.06) \quad (1.21)$$
$$\qquad\qquad t=80.4 \quad t=1.9$$

where the dummy D8 connotes groups of size eight (the n=4 groups are the omitted category). This 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

[11] See Mahadeva and Sterne (2000).
[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).

play, that is, in Part One of the experiment. We call this variable A_i (for ability) and use both it and its square as controls for skill in the following regression:

$$(5) \quad S_i = -300.5 + 1.29\,D8_i + 9.63A_i - 0.060A^2_i \qquad R^2 = 0.235, \quad N=504$$
$$\underset{t=2.4}{(124.1)} \quad \underset{t=1.8}{(0.72)} \quad \underset{t=2.9}{(3.28)} \quad \underset{t=2.8}{(0.022)}$$

Notice the huge jump in R^2—the variable A has high explanatory power.[13]

This regression 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%—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.

The 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 One 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, largely an artifact of the inflexible quadratic functional form. If 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. That said, it is still surprising that groups reap no further rewards from the individual abilities of their members once A exceeds a modest level (approximately 80). But this is a pretty robust finding that survives experiments with several functional forms.

[13] When (5) is estimated by ordinary least squares instead, the coefficients are almost identical, but the standard errors are roughly half of those in (5)—indicating that clustering matters.

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.[14] 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 in groups of four than in groups of eight. To see whether that factor might be empirically important in these data, we included both the average score of the group's best player (BEST) and its square in the regression to get:

$$(6) \quad S_i = -293.2 + 1.03\,D8_i + 7.03A_i - 0.044A^2_i + 2.02BEST_i - 0.010BEST^2_i$$

	(85.6)	(0.65)	(2.42)	(0.016)	(1.86)	(0.012)
	t=3.4	t=1.6	t=2.9	t=2.7	t=1.1	t=0.9

$R^2 = 0.261$ N = 504

The effect of larger group size is reduced by another 20%, to just one point, and it is now no longer significant at even the 10% 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 to those of the A variables. Moreover, adding BEST and $BEST^2$ raises R^2 by only 0.026.[15] However, an F-test of the joint hypothesis that the coefficients on both variables are zero strongly rejects that hypothesis (F=30.9, p = 0.00).[16] Thus, the evidence suggests that the fuller specification (6) is preferred, but that the influence of the

[14] 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).

[15] 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.

[16] 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--by 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.

best player on group decisionmaking is modest—a point to which we shall return in considering the effects of leadership.

Next, we consider whether heterogeneity of the members of the group, as measured by skill differences across players, improves group performance Specifically, we measure heterogeneity by introducing the variable SD_i, which is the standard deviation of the average scores obtained by the members of the group in Part One.[17] Adding this variable to regression (6) yields:

$$(7) \quad S_i = -293.4 + 1.03\,D8_i + 7.08A_i - 0.04A^2_i + 1.98BEST_i - 0.01BEST^2_i$$
$$(86.6) \quad (0.66) \quad (2.63) \quad (0.02) \quad (1.90) \quad (0.012)$$
$$t=3.4 \quad t=1.6 \quad t=2.7 \quad t=2.6 \quad t=1.0 \quad t=0.9$$

$$+ 0.02SD_i \qquad R^2 = 0.261 \; N = 504$$
$$(0.16)$$
$$t=0.1$$

Apart from the totally insignificant coefficient on SD, regression (7) looks almost exactly like regression (6). Thus heterogeneity does not seem to matter.

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 see what gives larger groups their small edge, we next examine 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. This was the variable that held the biggest surprise in our previous research: Groups actually had shorter mean LAGs than individuals, although the difference was not statistically significant.

[17] 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.

To determine whether a shorter or longer decisionmaking lag is the source of the advantage for large groups, we regress LAG on a dummy for the size of the group and the ability controls mentioned above, clustering by session as usual. The result is:

$$(8) \quad LAG_i = 97.3 - 0.02\, D8_i - 2.33A_i + 0.014A^2_i \quad R^2 = 0.066 \; N = 504$$
$$(33.7) \quad (0.42) \quad (0.91) \quad (0.006)$$
$$t=2.9 \quad t=0.1 \quad t=2.6 \quad t=2.4$$

This regression indicates no difference between the two group sizes in terms of speed of decisionmaking. (The p value of the coefficient of the dummy is 0.58.) Differences in ability are again significant, with groups comprised of more skilled players tending to decide more quickly—but only until A reaches 81.2. Moreover, the low R^2 in this regression indicates that neither group size nor ability explains much of the variation in lag times.

Next, we turn to *accuracy* rather than *speed*. Define the variable CORRECT to be equal to 1 if the group's initial interest rate move is in the correct direction—that is, a rise in G is followed by a monetary tightening, or a decline in G is followed by a monetary easing—and to be 0 otherwise. Do larger groups derive their advantage by being more accurate, in this sense?

Using the same right-hand side variables as in (8), we obtain:[18]

$$(9) \quad CORRECT_i = 0.44 - 0.01\, D8_i + 0.006A_i + 0.000A^2_i$$
$$(4.26) \quad (0.04) \quad (0.114) \quad (0.001)$$
$$t=0.1 \quad t=0.3 \quad t=0.05 \quad t=0$$
$$R^2 = 0.008 \quad N = 504$$

Once again there is no difference between groups of size four and size eight. It is interesting to note that the average ability of the members of the group is also of no use in

[18] 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.

predicting the group's odds of making the first interest rate move in the correct direction—a surprising finding.

Having failed so far, we turn finally 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 variable FREQ, which measures the number of rate changes a group makes over the course of a 20-quarter game. Since interest rate changes are costly, it pays for groups to economize on them. The usual simple regression reveals a modest effect of group interaction in producing more "patient" decisionmaking:

$$\underset{(10)}{FREQ_i} = \underset{(13.6)\ t=0.4}{6.07} - \underset{(0.15)\ t=1.8}{0.27\,D8_i} - \underset{(0.37)\ t=0.4}{0.13A_i} + \underset{(0.002)\ t=0.4}{0.001A^2_i}$$

$R^2 = 0.031$ N = 504

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 0.26 fewer interest rate changes per game. Since only about 2.25 changes are made on average, this is a meaningful difference, with a p-value of 0.08.

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" decisionmakers than individuals. This slightly more patient behavior, in turn, produces a systematic, though quite modest, performance advantage over small groups.

IV. Does leadership enhance group performance?

Up to now, we have focused on group size while ignoring the effects of leadership on performance. But as noted in the introduction, virtually all decisionmaking 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 in some sense. 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 begin with a simple regression comparing the scores (S) of groups with and without leaders—ignoring, for the moment, group size. Defining a dummy LED to be 1 if the group has a designated leader and 0 otherwise and controlling for ability, a regression over all 504 group observations yields:

$$(11) \quad S_i = -325.4 - 0.16\,LED_i + 10.30A_i - 0.064A^2_i$$
$$\qquad\quad (133.6) \quad (0.74) \qquad (3.51) \quad (0.023)$$
$$\qquad\quad t=2.4 \quad t=0.2 \qquad t=2.9 \quad t=2.8$$
$$R^2 = 0.227 \qquad N = 504$$

The effects of ability on group performance resemble regression (5), with a quadratic in A that peaks at 80.4. Of greater interest, however, is the regression coefficient on leadership. Regression (11) indicates a small *negative* effect of leadership (under 1 point), but it does not come close to statistical significance. 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 top scores during Part One purely by chance, and thus are not really any better at playing the game than the others. This possibility, however, is easily dismissed by looking at scores in Part Three—when subjects play again as individuals. Across all individuals who participated in the sessions with designated leaders, the correlation between Part One scores and Part Three scores is 0.45, indicating substantial, and durable, individual effects. Thus it was not just luck; some people do play the game better.

One interesting question to ask 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 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 Part One score of the highest-scoring individual—the very person who becomes the designated the leader in Part Two. So we run the following horse-race regression:

$$(12) \quad S_i = -393.6 + 12.26A_i - 0.078A^2_i - 0.38BEST_i + 0.005BEST^2_i$$
$$(202.2) \quad (6.10) \quad (0.041) \quad (2.70) \quad (0.017)$$
$$t=1.9 \quad t=2.0 \quad t=1.9 \quad t=0.1 \quad t=0.3$$
$$R^2=.322 \qquad N = 264$$

Interestingly, the average skill of the group's members is a much better predictor of performance than the skill of the leader. To see this formally, we ran F-tests to determine the effect of omitting the two A_i variables versus omitting the two $BEST_i$ variables from the regression. For the A_i variables, the F-statistic is 8.7 ($p = 0.00$) whereas for the $BEST_i$ variables, the F-statistic is only 3.2 ($p = 0.06$). The comparative weakness of the BEST variable helps to explain the absence of any leadership effects on performance: While the

leader is the best player, he or she seems incapable of improving the performance of the group.[19]

We next ask whether leadership effects on group performance differ by the gender of the leader, controlling for the group's ability, by adding the dummy variable FEMALE to the regression. Again, we restrict our attention to sessions with designated leaders:[20]

(13) $S_i = -740.63 + 21.33A_i - 0.137A^2_i - 0.63FEMALE_i$
(133.11) (3.61) (0.024) (1.05)
t=5.6 t=5.9 t=5.6 t=0.6
R^2=.368 N = 216

While the regression indicates a negative coefficient for female leaders, the magnitude of the coefficient is quite modest and it does not come close to statistical significance. Thus, women do neither better nor worse as leaders.[21]

So leaders seem to have no discernible effect on the quality of a group's overall performance. Do they, however, influence the group's strategy? To examine this, we look first at the dependent variable LAG defined earlier. Regression (14) shows that leadership does not influence the speed of reaction significantly:

(14) $LAG_i = 99.3 - 0.29\,LED_i - 2.38A_i + 0.015A^2_i$
(30.3) (0.41) (0.82) (0.006)
t=3.3 t=0.7 t=2.9 t=2.6)
R^2 = 0.068 N = 504

The coefficient of LED is negative, but insignificant.

[19] The inverted quadratic in BEST 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.
[20] A leader in one of the eight person sessions refused to identify his or her gender, which reduced the number of observations to 216.
[21] They are also neither better nor worse as followers. The sex composition of the group does not help explain the group's performance.

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

(15) $CORRECT_i = 0.35 - 0.025\ LED_i + 0.009A_i + 0.000A^2_i$

(3.82) (0.033) (0.102) (0.001)

t=0.1 t=0.7 t=0.1 t=0.03

$R^2 = 0.010$ N = 504

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 following regression shows. But the effect does not come close to statistical significance.

(16) $FREQ_i = 10.6 + 0.15\ LED - 0.26A_i + 0.002A^2_i$

(13.0) (0.15) (0.35) (0.002)

t=0.8 t=1.0 t=0.8 t=0.8

$R^2 = 0.019$ N = 504

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 next regression includes dummies for both group size and for leadership, allowing an interaction between the two:

(17) $S_i = 87.05 - 3.01\ LED_i - 0.002D8_i + 4.35(D8_i * LED_i)$

(0.72) (1.96) (1.23) (2.27)

t=121.4 t=1.5 t=0.0 t=1.9

$R^2 = 0.057$ N = 504

Here we find a surprisingly strong interaction effect, with a p-value of 0.06. Leadership actually hurts performance in groups of four (though the p-value of the negative

coefficient is only 0.13), but helps in groups of eight. Put differently, larger groups appear to do better if they are led, but smaller groups do worse.

Unfortunately, this effect is largely an illusion attributable to the fact that the {n=8, leader} groups just happened to draw better-than-average participants while the {n=4, leader} groups happened to draw some of the worst. This fact is shown in Table 3, and its implications are shown in regression (18), which augments (17) by controlling for ability in the usual way.

Table 3
Average Scores in Part One, by Treatment

Treatment	Part One Mean Score (individual play)
All treatments	77.4
n=4, no leader	78.4
n=4, leader	75.5
n=8, no leader	76.8
n=8, leader	78.2

$$(18)\quad S_i = -292.0 - 0.72\,LED_i + 0.77D8_i + 1.05(D8_i * LED_i) + 9.43A_i - 0.06A^2_i \qquad R^2 = 0.237 \quad N = 504$$

(121.0) (1.10) (0.84) (1.44) (3.18) (0.02)
t=2.41 t=0.7 t=0.9 t=0.7 t=3.0 t=2.4

This regression reveals that much of the difference in the performances of groups with and without leaders really reflects the different skill levels of the individual group members. For example, the coefficient on the interaction effect is reduced to less than one-fourth of its value in regression (17) and is now totally insignificant (p value=0.47). Still, the coefficients do suggest 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.

One other place to look for leadership effects is in how much people learn from their experience playing as a group. In our Princeton experiment (Blinder and Morgan (2005)), we found significant improvements in performance when individuals came together to play as groups. And the next section will show that the advantage for groups is even larger in the Berkeley experiment. Could it be that the learning that takes place during group play is greater when the group has a designated leader?

Table 4 displays the *improvements* in score from Part One (individual play) to Part Two (group play) separately for each of the four experimental treatments. While the individuals in the {n=4, leader} treatment groups stand out as the worst players in both parts, there is no support here 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 4: the average score of a given subject in group play (Part Two of the game) *minus* that individual's average score while playing as an individual in Part One.

Table 4: **Improvements from Individual to Group Play, by Treatment**

(1) ***Treatment***	***(2)*** ***Part One Mean Score (individual play)***	***(3)*** ***Part Two 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%)

Table 4 above suggests that improvements are systematically higher with larger groups

but independent of leadership. Thus, we include as righthand variables dummies for group size and whether the group was led or not. As usual, we cluster by session to obtain:

(19) $DIFF_i = 8.71 + 0.03\ LED_i + 1.46\ D8_i$ $\quad R^2 = 0.005 \quad N = 250$
(0.83) (0.99) (0.99)
t=10.5 t=0.03 t=1.5

This regression shows that leadership has no effect on the *improvement* between individual and group play. On the other hand, participation in larger groups improves upon individual performance slightly more than participation in smaller groups does; 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 our Princeton experiment (and replicate below) that scores typically improve quite a bit when subjects move from individual play to group play (from Part One to Part Two) but then fall back somewhat when they return to individual play (from Part Two to Part Three). The change in an individual's performance from Part One to Part Three 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 5 suggests that the answer is no. The subjects learn more from group play with a designated leader when n=4, but not when n=8. Notice, by the way, that the largest improvement in Table 5 comes in the {n=4, leader} groups, the very treatment that, by chance, got the weakest players. We will return to this point later.

Table 5
Improvements from Part One to Part Three, by Treatment

(1) *Treatment*	*(2)* *Part One Mean Score (individual play)*	*(3)* *Part Three 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%)

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 Three of the game less that individual's average score in Part One, on dummy variables for leadership and size. Clustering by session as usual, the result is:

(20) $POSTDIFF_i = 7.38 + 0.41\ LED_i - 0.18\ D8_i$
(1.13) (1.21) (1.21)
t=6.5 t=0.3 t=0.2
$R^2 = 0.001$ N = 250

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. If anything, the evidence points weakly in the other direction. 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.

Why no leadership effects?

First, in defense of our experimental design, note that we do *not* choose the leaders randomly or arbitrarily. Instead, 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—just as is normally the case in real-world groups. At least that was our intent. A second element of realism derives from the reward structure. By doubling the leader's reward in group play, we give him or her a greater stake in the outcome—just as leaders of real-world groups normally have a greater stake in the outcome than other members do. For example, history will appraise the performance of the "Greenspan Fed" and the "Rehnquist Court." The names of most of the other members will be forgotten.

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 decisionmaking when groups made decisions unanimously rather than by majority rule. And, as noted earlier, tie votes were rare.

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 might 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, p. 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. So the result may be relevant to actual monetary policy committees.

Fifth, however, there is never any disagreement among members of our ersatz MPCs over what the group's objectives (including the relative weights) are. 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 least on some real-world MPCs (e.g., the FOMC). Circumstances like that might allow more scope for effective leadership.

Sixth, and related, our committees deal only with "normal" monetary policy decisions. It is possible that greater scope for leadership might emerge if our experimental subjects were faced with crises.

Seventh, it might just be that the optimal committee size is, say n=6. In that case, committees of four (too small) and eight (too large) might be (approximately) 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 bigger differences.

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, for example, in the admittedly chauvinistic statement, "He's a leader of men." Our experimental

[22] This possibility was suggested to us by Petra Geraats, noting that Sibert (2006) suggested that the optimal committee size is five.

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, *et al.* (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 our 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 that 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, 90-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 we invite other researchers to pick up the challenge.

V. Groups versus individuals

We turn now, albeit very briefly, to the data 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.[23]

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%, 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%—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. With the usual correction for robust standard errors, we estimate:

$$(21)\quad S_i = 85.27 + 3.02\,GP_i - 4.18\,BERK_i + 2.50\,(GP_i * BERK_i)$$
$$\qquad\quad (0.37)\quad (0.57)\qquad (0.55)\qquad (0.75)$$
$$\qquad\quad t=231.8\quad t=5.4\qquad t=7.6\qquad t=3.4$$
$$R^2 = 0.027 \qquad N = 8{,}893$$

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

[23] 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 merging the two samples.

Berkeley students perform worse than Princeton students when playing as individuals, but improve more than Princeton students from group interaction. We do not have a ready explanation for this difference, but we do note that Lombardelli *et al.* (2005, p. 194) found that weaker players improved more over the course of their entire experiment—spanning both group and individual play.

This suggests a systematic pattern: that weaker players 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 One (individual play) to Part Two (group play) correlated *negatively* with the Part One scores. That is, do weaker players benefit more from working in groups? To examine this question, we regress the mean score of a group over its 12 repetitions (Gmean) on the average score of the individuals comprising the group while playing alone in Part One (A_i). The results are:

(22) $Gmean_i = 56.77 + 0.386\ A_i$ $\quad R^2=0.320 \quad N = 351$

(8.90) (0.11)

t=6.38 t=3.50

Notice that the coefficient on the average individual score is considerably smaller than one, which implies that $\partial(Gmean - A)/\partial A$ is decidedly negative (estimated to be -0.61). Thus, consistent with the findings of Lombardelli *et al*. (2005), we find that weaker players improve more from group interaction than do stronger players.

The next question pertains to the decisionmaking 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.

Formally, regression (23) estimates the same specification as (21), but with LAG replacing S as the dependent variable:

$$(23)\ LAG_i = 2.45 - 0.15\ GP_i + 0.75\ BERK_i + 0.12\ GP_i * BERK_i$$
$$(0.23)\quad (0.21)\quad (0.28)\quad (0.30)$$
$$t=10.7\quad t=0.7\quad t=2.7\quad t=0.4$$

$R^2 = 0.007 \qquad N = 8{,}893$

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 do Princeton students.

VI. 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.

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 purely 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 to supervisory positions 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.

References

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

Blinder, Alan S., *The Quiet Revolution: Central Banking Goes Modern,* Yale University Press, 2004.

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

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

Chappell, Henry W., Jr., Rob Roy McGregor, and Todd Vermilyea, *Committee Decisions on Monetary Policy,"* MIT Press, 2005.

Edmondson, A. "Psychological Safety and Learning Behavior in Work Teams." *Administrative Science Quarterly* 44 (4, December 1999): 350–383.

Fiedler, F. and F. Gibson, "Determinants of Effective Utilization of Leader Abilities," in *Concepts for Air Force Leadership,* R.I. Lester and A.G. Morton, eds., Air University Press, Melbourne, 2001.

Guth, Werner, M.Vittoria Levati, Matthias Sutter, and Eline van der Heijden, "Leadership and Cooperation in Public Goods Experiments," Discussion papers on strategic interaction no. 2004-29, Max Planck Institute of Economics, 2004.

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

Mahadeva, and Gabriel Sterne, eds., *Monetary Policy Frameworks in a Global Context*, Routledge Publishers, New York, 2000.

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

Rudebusch, Glenn, "Is the Fed Too Timid?: Monetary Policy in an Uncertain World," *Review of Economics and Statistics* 83(2, May 2001): 203–217.

Sibert, Anne, "Central Banking by Committee," *International Finance*, 9(2, August 2006): 145–168.

White, Halbert, “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity,” *Econometrica* 48(May 1980): 817-838.