Draft of October 18, 2006

Leadership in Groups: A Monetary Policy Experiment

by

Alan S. Blinder and **John Morgan**
Princeton University **University of California, Berkeley**

November 2006

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.

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. That experiment 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 not from 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 identical 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 neither the main purpose nor the focus of this paper. Instead, we study two

[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.

important issues that were deliberately omitted from our previous experimental design.

The first pertains to *group size*. In the Princeton experiment, every student monetary policy committee (MPC) had five members—precisely (and coincidentally) the size that Sibert (2005) 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 size four and size 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 known as 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.[3] Observed reality, therefore, strongly suggests that groups need leaders in order

[2] The reason for choosing even-numbered groups will be made clear shortly.
[3] Juries come close, but even they have a foreman.

to perform well. But is it true? That is the main question that this research is designed to answer.

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 (less so, it seems, under Ben Bernanke) was at one extreme; it 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, even famously allowed himself to be outvoted once in 2005 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 decisionmaking. Some of these committees are firmly led; others are led only gently.

The scholarly literature on group decisionmaking, which comes mostly from psychology and organizational behavior rather than from economics, gives us relatively little guidance on what to expect. And only a small portion of this literature is experimental. As a broad generalization, our quick review of the literature led us to expect to find some positive effect 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 the key ingredient in effective leadership (Fiedler and Gibson, 2001). Rather, the 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. 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 ten or eleven sessions of 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, though there are a few exceptions. Section VI summarizes the conclusions.

II. The Experimental Setup[4]

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 that they would be playing a monetary policy game. 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:

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

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

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

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 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, of course, any constant inflation rate can be an equilibrium.

We begin each round of play with an initial inflation rate of 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].[5] 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."

Lest our subjects had forgotten their basic macroeconomics, the instructions reminded them that raising the interest rate lowers inflation and raises unemployment, while lowering it does the reverse, albeit with a lag.[6] 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, however, told that the demand shock, whose magnitude they do not know, will occur at a random time that is equally likely to be any of periods 1 through 10.

While this model may look trivial, stabilizing such a system 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 long lags and modest early-period effects combine to make the divergence from equilibrium pretty gradual, and hence less than obvious at first.

[5] The distributions are iid and uniform, rather than normal, for programming convenience.
[6] A copy of the instructions is available on request.

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 normally close to the optimal 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 students, 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 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:

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

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 ready 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 for that period.[7] At the end of the session, scores are converted into money at the rate of 25 cents per percentage point. Subjects typically earned about $20-$21 out of a theoretical maximum of $25.

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.[8] 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 they must

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

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

add, say, a quadratic term in (i_t – i_{t-1}) to fit the data. Without that wrinkle, interest rates are far too volatile.[9]

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. They 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 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 telling everyone which student earned the highest score while playing alone in Part One, and designates that student as the "leader" (the term we used) of the group for Part Two. Up to that moment, the subjects do not know that leadership wioll have anything to do with the experiment. The leaders of each group is responsible for communicating (verbally) the group's decision to the

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

experimenter, who then enters it into the computer. That device normally ensures that the leader leads the discussion, or at least that communication is directed to him or her. The leader is accorded two privileges, which are common knowledge: his or her score in Part Two is *double* that of the other subjects; and, in the event of a tie vote, the leader gets to break the tie. That is why we chose even-numbered groups.[10]

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.

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

[10] In fact, ties were rare.

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 rightmost 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

III. Are larger groups more effective than smaller groups?

The title of our 2005 paper asked, "Are two heads better than one?" We now ask 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. Initially, we do not attempt to control for the skill

[11] See Mahadeva and Sterne (2000).

levels of the members of the group either. Regressing the average game score (the variable S defined above) for each of the 504 group observations on only 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]

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

where D8 is a dummy for groups of size eight (the n=4 groups are the omitted category). 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, it is possible that larger groups simply had, on average, more highly-skilled individuals than did smaller groups. Therefore, it might be important 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, 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:

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

[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 the huge jump in R^2—the variable A has real 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%. 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. This is an unexpected and puzzling finding which raises many questions; so it merits a brief digression.

First, is it a fluke? We went back to our Princeton data and discovered that the same general quadratic shape of the S=f(A) function held in those data, although neither A nor A^2 was statistically significant. Second, does the slope really turn down, rather than just flatten out, at high values of A? To study this, we ran several "horse race" regressions. One compared the fits of the quadratic functional form in (5) with that of a logarithmic specification; the quadratic fit the data better. The other nested the two specifications by including A, A^2, and ln(A) in the same regression and then doing F-tests; those tests found the two quadratic coefficients to be jointly significant while the coefficient on ln(A) was not. So the function does seem to turn down. Third, however, the estimated

[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.

negative effect of high *individual* ability on *group* performance is not quantitatively strong (in the relevant range). For example, while the slope ∂S/∂A is 1.27 when A=70, it is just -0.40 when A=84. In sum, this is a weak effect, but it does seem to be present in the data.

That said, let us return to why larger groups perform better than smaller groups. One possible explanation is that a group's decisions might be 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:

$$
\begin{array}{llccccccc}
(6) & S_i = & -293.2 & + \; 1.03\, D8_i & + \; 7.03A_i & - \; 0.044A^2_i & + \; 2.02BEST_i & - \; 0.010BEST^2_i \\
 & & (85.6) & (0.65) & (2.42) & (0.016) & (1.86) & (0.012) \\
 & & t=3.4 & t=1.6 & t=2.9 & t=2.7 & t=1.1 & t=0.9
\end{array}
$$

$R^2 = 0.261$ $N = 504$

The effect of larger group size is reduced by another 20%, to just one point, and it is now no longer significant at standard levels (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. However, an F-test of the joint hypothesis that the coefficients on both variables are zero strongly rejects that

[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).

hypothesis (F=30.91, p = 0.00). Thus, the evidence suggests that the fuller specification (6) is preferred, but that the influence of the 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 among the players, improves group performance. Specifically, we measure heterogeneity by introducing the variable SDA_i, which is the standard deviation of the average scores obtained by the members of the group in Part One. Adding this variable to regression (6) yields:

$$
\begin{array}{lllllll}
(7)\quad S_i = & -293.4 & +\ 1.03\ D8_i & +\ 7.08A_i & -\ 0.04A^2_i & +\ 1.98BEST_i & –\ 0.01BEST^2_i \\
& (86.6) & (0.66) & (2.63) & (0.02) & (1.90) & (0.012) \\
& t=3.4 & t=1.6 & t=2.7 & t=2.6 & t=1.0 & t=0.9 \\
& & & & & & \\
& +\ 0.02SDA_i & & R^2 = 0.261 \quad N = 504 & & & \\
& (0.16) & & & & & \\
& t=0.1 & & & & &
\end{array}
$$

Apart from the totally insignificant coefficient on SDA, 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: Why is this the case? To try to understand the source of the larger group's (slightly) superior performance, 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 LAGs, on average, than individuals, although the difference was not statistically significant.

To determine whether a shorter or longer decisionmaking lag is the source of the advantage for large groups, we start off by regressing LAG on a dummy for the size of the group, clustering by session as usual. The result is:

$$
\begin{array}{llllll}
(8) & LAG_i \;=\; & 3.31 & -\; 0.27\, D8_i & R^2 = 0.002 & N = 504 \\
& & (0.35) & (0.48) & & \\
& & t=9.5 & t=0.6 & &
\end{array}
$$

The regression indicates no significant difference between the two groups in terms of the speed of decisionmaking. (The p value of the coefficient of the dummy is 0.58.) Controlling for differences in ability, which are once again significant, reduces even this small negative coefficient (which means that larger groups decide faster) to essentially zero:

$$
\begin{array}{llllllll}
(9) & LAG_i \;=\; & 97.3 & -\; 0.02\, D8_i & -\; 2.33 A_i & +\; 0.014 A^2_i & R^2 = 0.066 & N = 504 \\
& & (33.7) & (0.42) & (0.91) & (0.006) & & \\
& & t=2.9 & t=0.1 & t=2.6 & t=2.4 & &
\end{array}
$$

Groups with more skilled players tend to decide more quickly, but only until A reaches 81.2. But there is no case at all that larger groups are faster. Moreover, the low R^2 values in these regressions indicate 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?

As usual, we start with the simplest specification:[15]

$$\begin{array}{llllll} (10) & \text{CORRECT}_i \;= & 0.889 & +\; 0.016\,\text{D8}_i & R^2 = 0.001 & N = 504 \\ & & (0.031) & (0.035) & & \\ & & t=28.9 & t=0.4 & & \end{array}$$

And once again, there is no difference between groups of size four and size eight. The next regression shows that controlling for skill levels does not change this conclusion:

$$\begin{array}{llllll} (11) & \text{CORRECT}_i \;= & 0.44 & -\; 0.01\,\text{D8}_i & +\; 0.006\text{A}_i & +\; 0.000\text{A}^2_i \\ & & (4.26) & (0.04) & (0.114) & (0.001) \\ & & t=0.1 & t=0.3 & t=0.05 & t=0 \end{array}$$

$R^2 = 0.008 \quad N = 504$

As before, group size has no effect. What is interesting to note here is that the average ability of the members of the group is also of no use in 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 test for this, we define a variable FREQ, which measures the average number of rate changes a group makes over the course of a 20-quarter game. Since interest rate changes

[15] 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.

are costly, it pays for groups to economize on them. The initial simple regression reveals a modest effect of group interaction in producing more “patient” decisionmaking:

$$\begin{array}{lllll} (12) & \mathrm{FREQ}_i \; = & 2.08 & - \;\; 0.26\,\mathrm{D8}_i & R^2 = 0.017 \quad N = 504 \\ & & (0.10) & (0.15) & \\ & & t=20.0 & t=1.7 & \end{array}$$

Here at last we find a partial answer to the question of why larger groups perform 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, although the p-value of the coefficient is only 0.10. But could it simply be that more skilled players manage to economize on rate changes better than less skilled players? The answer turns out to be no. It really is a large-group effect, albeit a modest one, as the following regression shows:

$$\begin{array}{llllll} (13) & \mathrm{FREQ}_i \; = & 6.07 & - \;\; 0.27\,\mathrm{D8}_i & - \;\; 0.13A_i & + \; 0.001A^2_i \\ & & (13.6) & (0.15) & (0.37) & (0.002) \\ & & t=0.4 & t=1.8 & t=0.4 & t=0.4) \\ \multicolumn{6}{l}{R^2 = 0.031 \quad N = 504} \end{array}$$

Indeed, strikingly, the ability variable seems to have little to do with the frequency of rate changes.

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 this slightly more “patient” behavior produces a systematic, though quite modest, performance improvement over small groups.

IV. Does leadership enhance group performance?

Up until 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 functional. 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, as usual, 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, a regression over all 504 group observations yields:

$$(14) \quad S_i = \underset{\substack{(0.61)\\ t=142.0}}{87.05} - \underset{\substack{(1.22)\\ t=0.7}}{0.83}\, LED_i \qquad R^2 = 0.004 \qquad N = 504$$

The regression coefficient indicates a small *negative* effect of leadership (under 1 point), but it does not come close to statistical significance; the p-value is almost exactly 0.5—a coin flip. The basic finding is that leadership does not affect group performance.

We proceed now to try to overturn this surprising non-result. First, could it be that a positive effect of leadership is masked because the participants in the sessions with leaders just happen to be, on average, less able than those in the sessions without leaders? Adding controls for ability (A and A^2) as we did before yields:

$$
\begin{array}{llllll}
(15) & S_i = & -325.4 & -\ 0.16\ LED_i & +\ 10.30A_i & -\ 0.064A^2_i \\
 & & (133.6) & (0.74) & (3.51) & (0.023) \\
 & & t=2.4 & t=0.2 & t=2.9 & t=2.8)
\end{array}
$$

$R^2 = 0.227$ N = 504

The estimates resemble regression (5), with a quadratic in A that peaks at 80.4. The estimated effect of leadership here is negative, but trivially so; essentially, it is zero.

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 score of the highest-scoring individual during Part One of the experiment. Since that person was designated as the leader in Part Two, BEST also measures the leader's ability. So we run the following horse-race regression:

$$
\begin{array}{lllllll}
(16) & S_i = & -393.6 & +\ 12.26A_i & -\ 0.078A^2_i & -\ 0.38BEST_i & +\ 0.005BEST^2_i \\
 & & (202.2) & (6.10) & (0.041) & (2.70) & (0.017) \\
 & & t=1.9 & t=2.0 & t=1.9 & t=0.1 & t=0.3
\end{array}
$$

R^2=.322 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.[16]

Similarly, we can ask whether leadership effects on group performance differ by the gender of the leader, controlling for ability. Again, we restrict our attention to sessions with designated leaders and add the dummy variable FEMALE to the regression.[17]

$$\begin{array}{llllll} (17) & S_i \;= & -740.63 & +\; 21.33A_i & -\; 0.137A^2_i & -\; 0.63FEMALE_i \\ & & (133.11) & (3.61) & (0.024) & (1.05) \\ & & t=5.6 & t=5.9 & t=5.6 & t=0.6 \\ R^2=.368 & N = 216 & & & & \end{array}$$

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.[18]

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 (18) shows that leadership does not influence the speed of reaction significantly.

$$\begin{array}{lllllll} (18) & LAG_i \;= & 3.24 & -\; 0.11\,LED_i & & R^2 = 0.000 & N = 504 \\ & & (0.37) & (0.49) & & & \\ & & t=8.7 & t=0.2 & & & \end{array}$$

[16] The inverted quadratic in BEST looks peculiar, but it is upward-sloping in the relevant range.

[17] A leader in one of the eight person sessions refused to identify his or her gender; hence the number of observations is reduced to 216.

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

A slightly negative coefficient (indicating shorter lags) also appears when we control for the group's ability in regression (19) below. But it, too, is insignificant.

$$\begin{array}{lllllll}
(19) & LAG_i = & 99.3 & -\ 0.29\ LED_i & -\ 2.38A_i & +\ 0.015A^2_i \\
 & & (30.3) & (0.41) & (0.82) & (0.006) \\
 & & t=3.3 & t=0.7 & t=2.9 & t=2.6)
\end{array}$$

$R^2 = 0.068$ $N = 504$

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

$$\begin{array}{llllll}
(20) & CORRECT_i = & 0.913 & -\ 0.030\ LED & R^2 = 0.002 & N = 504 \\
 & & (0.016) & (0.034) & & \\
 & & t=55.9 & t=0.9 & &
\end{array}$$

And, once again, controlling for skill levels does not change this conclusion:

$$\begin{array}{llllll}
(21) & CORRECT_i = & 0.35 & -\ 0.025\ LED_i & +\ 0.009A_i & +\ 0.000A^2_i \\
 & & (3.82) & (0.033) & (0.102) & (0.001) \\
 & & t=0.1 & t=0.7 & t=0.1 & t=0.03
\end{array}$$

$R^2 = 0.010$ $N = 504$

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 two regressions show. But in neither case is the effect close to statistical significance.

$$\begin{array}{llllll}
(22) & FREQ_i = & 1.88 & +\ 0.14\ LED_i & R^2 = 0.005 & N = 504 \\
 & & (0.12) & (0.16) & & \\
 & & t=16.1 & t=0.9 & &
\end{array}$$

$$\begin{array}{llllll}
(23) & FREQ_i = & 10.6 & +\ 0.15\ LED & -\ 0.26A_i & +\ 0.002A^2_i \\
 & & (13.0) & (0.15) & (0.35) & (0.002) \\
 & & t=0.8 & t=1.0 & t=0.8 & t=0.8
\end{array}$$

$R^2 = 0.019$ $N = 504$

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 essentially combines (4) (for group size) and (14) (for leadership), allowing for an interaction between the two:

$$(24)\quad \begin{array}{lllll} S_i = & 87.05 & -\ 3.01\ LED_i & -\ 0.002D8_i & +\ 4.35(D8_i * LED_i) \\ & (0.72) & (1.96) & (1.23) & (2.27) \\ & t=121.4 & t=1.5 & t=0.0 & t=1.9 \end{array}$$

$R^2 = 0.057$ $N = 504$

In fact, we find a surprisingly strong interaction effect (with p-value=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 get better-than-average participants while the {n=4, leader} groups happened to get some of the worst. This fact is shown in Table 3, and its implications are shown in regression (25), which augments (24) 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

$$(25)\quad S_i = -292.0 - 0.72\ LED_i + 0.77D8_i + 1.05(D8_i * LED_i) +$$

(121.0) (1.10) (0.84) (1.44)
t=2.41 t=0.7 t=0.9 t=0.7

$9.43A_i - 0.06A^2_i$ $R^2 = 0.237$ N = 504
(3.18) (0.02)
t=3.0 t=2.4

This regression reveals that much of the difference in performance 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 (24) and is now totally insignificant (p value=0.47). Still, the coefficients 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 the experience of 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 apparently takes place in 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 Part One, there is no support here for the idea that group interactions help subjects more when there is a designated leader.

Table 4
Improvements from Individual to Group Play, by Treatment

(1)	***(2)***	***(3)***	***(4)***
Treatment	***Part One Mean Score***	***Part Two Mean Score***	***Difference***
	(individual play)	***(group play)***	
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%)

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 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:

$$(26)\quad DIFF_i = \underset{\substack{(0.83)\\ t=10.5}}{8.71} + \underset{\substack{(0.99)\\ t=0.03}}{0.03}\, LED_i + \underset{\substack{(0.99)\\ t=1.5}}{1.46}\, D8_i \qquad R^2 = 0.005 \quad N = 250$$

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 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 with leadership than without?

Table 5 suggests that the answer is no. When n=4, the subjects learn more from group play when the groups have a designated leader, 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*** ***(group 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 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:

$$\begin{array}{llll} (27)\ \ \mathrm{POSTDIFF}_i = & 7.38\ + & 0.41\ \mathrm{LED}_i\ - & 0.18\ \mathrm{D8}_i \\ & (1.13) & (1.21) & (1.21) \\ & t=6.5 & t=0.3 & t=0.2 \end{array}$$

$R^2 = 0.001$ $N = 250$

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 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 leaders perform no better 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, gives them a certain legitimacy—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, it should be noted that while giving the leader the tie-breaking vote should allow 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.

Third, 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 performance." That said, making monetary policy decisions is, for the most part, an intellective task. So the result may have relevance to actual monetary policy committees.

Fourth, 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, at least on some real-world MPCs (e.g., the Fed), there is potential for disagreement over the central bank's objectives and/or weight. In such cases, the leader might (or might not) be more influential.

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 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 (2005) 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 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. That said, the recent selection of an outstanding academic economist, Ben Bernanke, to be Chairman of the Federal Reserve Board suggests that intellectual ability is a key consideration in selecting leaders of real-world MPCs.

V. Groups versus individuals

We turn now 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 had found earlier with the Princeton sample. Still, a few differences are worth noting.

To begin with, we found in our original experiment that groups (which were then of size five) turned in better average performances than did individuals. Specifically, the average group score (on the 0-100 scale) 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).

Before commenting on Princeton-Berkeley differences, a few words on the two samples are in order. Table 6 displays summary statistics comparing individual scores at Berkeley and at Princeton. It shows that Princeton students performed notably better than Berkeley students—scoring about 5.8% higher when playing as individuals. We are not particularly interested in appraising the relative quality of Princeton versus Berkeley undergraduates, but this difference does align with performance on standardized tests for entering students at the two universities. In addition, notice that the standard deviation across the Berkeley scores is considerably higher than it was at Princeton. Thus there is considerably more variability in the performance of student monetary policymakers at Berkeley than was the case at Princeton. Indeed, the differences between Berkeley and Princeton subjects are statistically significantly different from one another. For example, if one compares mean individual performance in Part One of the experiment (prior to any exposure to group play), and treats each subject as the unit of observation, one obtains a mean for Berkeley students of 77.4 versus a mean of 83.9 for Princeton students. This large performance gap is highly significant (t = 5.8).

Table 6
Individual Scores: Berkeley and Princeton

	Berkeley	**Princeton**
Mean Score	81.1	85.3
Standard Deviation	15.8	10.1

Another noteworthy difference between the Princeton and Berkeley samples is that women's performance as individuals in Part One is worse than that of men in the Berkeley sample, but not in the Princeton sample. This can be seen in Table 7.

Table 7
Individual Scores Prior to Group Play, by Gender: Berkeley vs. Princeton

		Berkeley	**Princeton**
Females	Mean Score	75.94	83.37
	Standard Deviation	20.04	11.40
Males	Mean Score	79.38	84.38
	Standard Deviation	16.88	11.59

The 3.44 point gap between the scores of males and females at Berkeley is highly significant (t = 4.48, p = 0.00), whereas the 1.01 point gap for the Princeton subjects is not (t = 1.37, p = 0.18).

Interestingly, however, the inferior performance of women at Berkeley disappears once they participate in group decisionmaking. Table 8 presents the mean scores and their standard deviation, arrayed by gender and university, for individual play in Part Three, which comes *after* exposure to group play. The gender gap between the mean scores, which was substantial for Berkeley subjects in Part One of the experiment (see Table 7), virtually vanishes after exposure to group decisionmaking (see Table 8). For the Princeton subjects, the small initial gap in scores (favoring men) seen in Table 7 drops to essentially zero in Table 8. Naturally, neither of these gender differences comes close to statistical significance. Notice that these results imply that women learn more than men from group play.

Table 8
Individual Scores Following Group Play, by Gender: Berkeley vs. Princeton

		Berkeley	**Princeton**
Females	Mean Score	84.80	86.67
	Standard Deviation	9.49	7.75
Males	Mean Score	85.22	86.60
	Standard Deviation	10.63	8.52

Table 9 contains a parallel comparison of *group* scores at Berkeley and Princeton. Although the Berkeley scores come from groups of size four and size eight while the Princeton scores come from groups of size five, our earlier result that group size barely matters suggests that the comparison is valid. The group scores, like the individual scores, are higher at Princeton than at Berkeley. But notice that the difference in performance across the two experiments is now only about 2.1%--far lower than the percentage difference for individual play. This means that the performance gain from group play is higher for Berkeley students than for Princeton students.

Table 9
Group Scores: Berkeley and Princeton

	Berkeley	Princeton
Mean Score	86.6	88.3
Standard Deviation	6.8	4.7

The following regression confirms that this difference is significant, after the usual correction for robust standard errors. We estimate:

$$\begin{array}{lllll} (28)\ \ S_i = & 85.27 & +\ 3.02\ GP_i & -\ 4.18\ BERK_i & +\ 2.50\ (GP_i * BERK_i) \\ & (0.37) & (0.57) & (0.55) & (0.75) \\ & t=231.8 & t=5.4 & t=7.6 & t=3.4 \end{array}$$

$R^2 = 0.027 \qquad N = 8,893$

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 Berkeley students performed far worse than Princeton students when playing as individuals, but improved 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. Remember also that Berkeley women, who were weaker players than Berkeley men in Part One of the experiment, also improved more after group play.

All this suggests a systematic pattern: that weaker players gain more from exposure to group play. To investigate this phenomenon 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 (Smean) on the average score of individuals comprising the group in Part One of the game (A_i). The results are:

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

Notice that the coefficient on the average individual score is considerably smaller than one, which implies that $\partial(\text{Smean} - A)/\partial A$ is decidedly negative (in fact, it is 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 the difference was not statistically significant. 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 (30) estimates the same specification as (28), but with LAG replacing S as the dependent variable:

$$\begin{array}{lllll} (30)\ LAG_i = & 2.45 & -\ 0.15\ GP_i & +\ 0.75\ BERK_i & +\ 0.12\ GP_i\ {*}BERK_i \\ & (0.23) & (0.21) & (0.28) & (0.30) \\ & t=10.7 & t=0.7 & t=2.7 & t=0.4 \end{array}$$

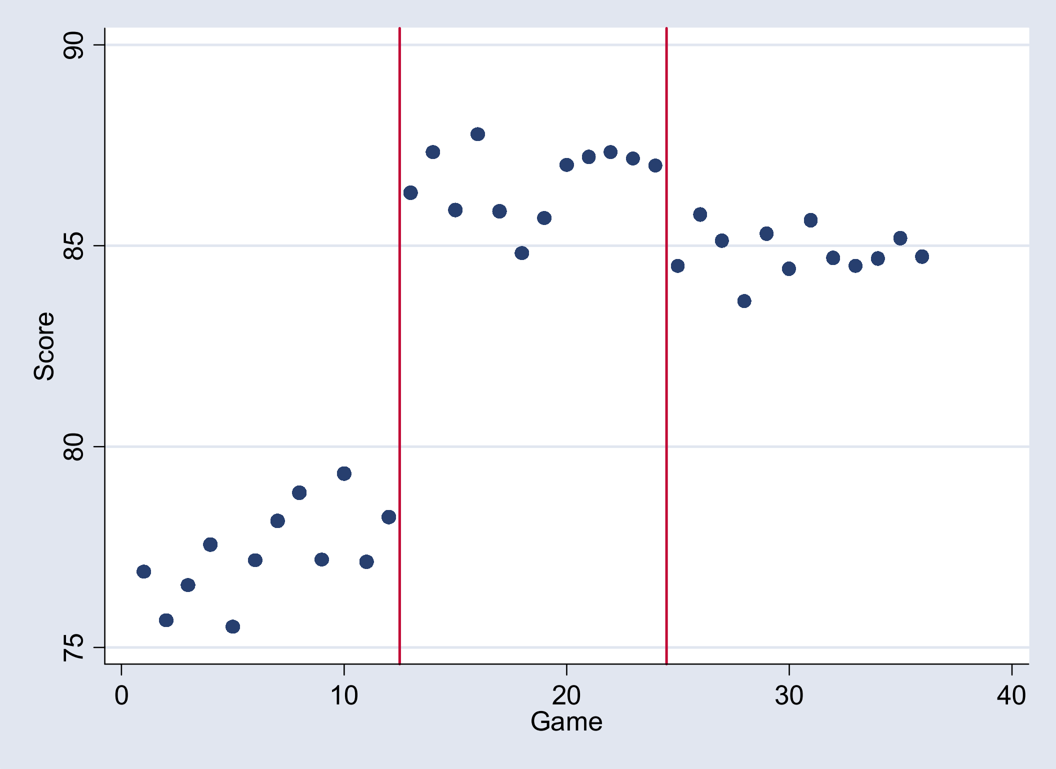
$R^2 = 0.007$ N = 8,893

The 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.

Finally, a few words on learning are in order. In our Princeton experiment, we found little evidence for conventional learning by doing, but strong evidence that subjects learned a great deal from their experience in group play.[19] Figure 1 displays parallel results for the Berkeley experiment. Visually, there is now a meaningful sign of improving scores during the 12 rounds of Part One (individual play)—which is once again consistent with the finding that weaker players learn more. But there is apparently no learning during the 12 rounds of Part Two (group play) or during the 12 rounds of Part Three (individual play again). Most notably, scores rise sharply when students move from individual to group play (from Part One to Part Two), and then fall back a bit when they return to individual play (from Part Two to Part Three), just as they did in our Princeton experiment.

[19] By contrast, Lombardelli *et a*l. (2005) found that subjects did learn from experience.

Figure 1: Scores over Time



To assess the statistical significance of these results, we regress S_i on a time trend whose slope and intercept are allowed to vary in each of the three periods of the game. Specifically, define the variable GAME to be a linear time trend that runs from 1 to 36, and the variables P2 and P3 to be dummy variables connoting Parts Two and Three of the experiment, respectively. We restrict attention to data from the Berkeley experiment (shown in Figure 1) and cluster standard errors by session. The results are:

$$\text{(31)}\quad \begin{array}{llllll} S_i = & 76.06 & + \; 0.20\,\text{GAME}_i & + \; 9.22\text{P2}_i & - \; 0.14(\text{P2}_i * \text{GAME}_i) & + \\ & (1.13) & (0.11) & (1.90) & (0.14) & \\ & t=67.5 & t=1.8 & t=5.0 & t=1.0 & \end{array}$$

$$\begin{array}{llll} 9.00\text{P3}_i & - \; 0.21(\text{P3}_i * \text{GAME}_i) & R^2 = 0.065 & N = 6{,}493 \\ (1.73) & (0.12) & & \\ t=5.2 & t=1.7 & & \end{array}$$

Notice that the time trend is positive and significant (p value = 0.075) in the first part of the game (an estimated gain of 0.2 point per round), whereas afterward, the time trends are pretty near zero—just as the graph suggests. The three intercept and three slope terms together indicate that moving from individual play in Part One to group play in Part Two has a strong and highly significant positive effect on scores (about 7.6 points[20]), about a quarter of which is lost when subjects return to individual play in Part Three.[21]

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. Furthermore, committees do not take any longer (as measured by required data inflow) to reach decisions. 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 shreds of suggestive (but not particularly surprising) 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

[20] The intercept jumps upward by 9.22, but the dummy P2 also turns on when GAME rises from 12 to 13.

[21] When GAME rises from 24 to 25, the intercept falls by 0.22, but the dummy P3 turns on while the dummy P2 turns off for the time trend terms. The net effect of all this, plus incrementing GAME by 1, is a reduction of 1.9 points.

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 ersatz 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 tasks in the economic world, including monetary policy, that are more intellective than combative in nature. 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 be a better institutional design than a *collegial* committee, where the role of the leader is more pronounced.

Finally, we unearth a small puzzle that cries out for explanation. We find that the decisionmaking ability of an experimental monetary policy committee is *not* monotonically increasing in the average ability of its members. Rather, the functional relationship rises to a peak (at above-average ability) and then falls (albeit only slightly)—perhaps mimicking the functionality of a university faculty meeting. Since there are no issues of either congestion or conflicting incentives in our experimental design, we are at a loss to explain this surprising finding. And, at this point, we are certainly not prepared to recommend that governments selecting candidates for real-world monetary policy committees adopt the Hruska principle—by adding a dash of mediocrity.[22]

[22] Roman Hruska was the U.S. Senator who defended an ill-fated nominee to the U.S. Supreme Court in 1970 by arguing that there are lots of mediocre people, and that they are entitled to representation on the Court, too!

References

Blades, J. W. “Influence of Intelligence”; 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, October 2005, pp. 789-812.

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, no. 4 (December 1999): 350-383.

Fiedler, F. and F. Gibson, “” in *Concepts for Air Force Leadership,* R.I. Lester and A.G. Morton, eds., Air University Press, Melbourne, 2001

Guth, W., M.V. Levati, M. Sutter, and E. van der Heiden “Leadership and cooperation in public goods experiments,” Discussion papers on strategic interaction No, 2004

Lombardelli, Clare, James Proudman, and James Talbot, Committees versus Individuals: An Experimental Analysis of Monetary Policy Decision Making,” International Journal of Central Banking, vol. 1, no. 1, June 2005, pp. 181-205.

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

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

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

Sibert, Anne, “Central Banking by Committee,” prepared for workshop at De Nederlandsche Bank, November 28, 2005.

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